

Proceedings of the 1st Symposium on AI and Cognitive Sciences in Education (AICSED) 2026

About the Symposium

Recent advances in artificial intelligence (AI) are transforming educational practice and raising both practical and theoretical questions about how learning, cognition, and agency develop in technology-rich environments. This symposium approaches education through an expanded conception of mind, informed by 4E cognition and predictive processing, and conceptualizes classrooms as cognitive niches in which tools, materials, and computational systems function as integral components of the learning process. From this perspective, educational environments do not merely deliver content; they systematically shape and reorganise learners' cognitive, motivational, metacognitive, and affective dynamics over development.

This symposium invited contributions that develop and assess intelligent tutoring systems, adaptive and multimodal interfaces, and assistive technologies that scaffold attention, curiosity, self-regulation, self-efficacy, and critical reasoning, while supporting collaborative knowledge construction and productive, error-tolerant inquiry. Particular emphasis has been placed on technically grounded approaches that integrate system architecture, instructional design, and empirical validation, and that demonstrate measurable gains in learner agency, accessibility, and learning outcomes through robust metrics.

AICSED 2026 formed part of the annual Convention of *The Society for the Study of Artificial Intelligence and Simulation of Behaviour* (AISB) 2026, which took place at the University of Sussex in England, UK, on 1-2 July 2026.

<https://aisb.org.uk/>

<https://aicsed.github.io/2026>

Program Committee

Symposium Chair:

Dr Ronald Grau, University of Sussex, England, UK

Co-Chairs:

Dr Grecia Garcia Garcia, University of Sussex, England, UK

Dr Daniel Hajas, University College London, UK

Dr Laura Desiree Di Paolo, University of Sussex, England, UK

Symposium talks

Keynote talks

Benedict du Boulay: “AI in Education – the first 60 years”	3
Jonathan Zong: “The Web Is Being Made Accessible for AI, Not People”	3

Part I - Human Learning, Pedagogy, and Design

Ronnie Videla, Dor Abrahamson and Marcelo Chávez: Embodied AI: Designing Inclusive Digital Ecologies.....	4
Grecia Garcia Garcia and Peter C-H. Cheng: Can an LLM build an application to automatically assess students’ interpretations of diagrams?	12
Daniel Hajas: Flow-Centred Design: Rethinking Accessibility as an Experiential Quality in Human- Computer Interaction.....	15
Christopher Luu and Ronald R. Grau: First steps towards novel adaptive systems for individualised music learning and assessment: An investigation of memory chunking in a piano chord production task.....	23
Dmitri Droujkov and Sutapa Dey Tithi and Tiffany Barnes: Confusion, Agency, and Metacognitive Instruction in an Intelligent Tutoring System: An Ability-Moderated Account	28

Part II - Technical and Systemic Aspects

Jonathan Zong, Isabella Pedraza Pineros, Mengzhu (Katie) Chen, Daniel Hajas and Arvind Satyanarayan: Semantic Scaffolding: Augmenting Textual Structures with Domain-Specific Groupings for Accessible Data Exploration.....	36
Charlotte Richardson and Ronald R. Grau: The Development and Evaluation of a Hybrid ViT-LSTM for Word-Level Sign Language Recognition.....	44
Alma Jazmín Castillo Rodríguez and Felipe De la Peña: Democratising Personalised STEAM Education: A Generative AI-Driven Intelligent Tutoring System Aligned with Standardised Pedagogical Frameworks in Mexico.....	50

OPENING KEYNOTE TALK

“AI in Education – the first 60 years”

This talk provides a brief overview of nearly 60 years of research in AI in education summarising both its positive and negative effects on learning. As part of this, it looks at the generations of AI technology from GOFAI to LLMs and their varying roles in education while emphasising the central importance and agency of human teachers in AI augmented classes. The talk concludes with a focus on the increasing importance of supporting and developing students’ self-regulated learning skills.

Speaker biography:

Benedict du Boulay is an Emeritus Professor of Artificial Intelligence at the University of Sussex. He gained a Ph.D. from the University of Edinburgh (1978). He worked in industry and as a schoolteacher prior to his PhD. His academic career continued at the Universities of Aberdeen and Sussex. His primary research area is the application of Artificial Intelligence in Education, particularly issues around modelling and developing students’ metacognition and motivation. He was President (2015-2017) of the International Society for Artificial Intelligence in Education. He co-edited the recent Handbook of AI in Education. A second edition is due out in 2027.

CLOSING KEYNOTE TALK

“The Web Is Being Made Accessible for AI, Not People”

When the Svelte framework added plaintext documentation for AI agents—“Beep boop”—it joined a wave of efforts to make the web legible to machines: llms.txt, MCP servers, LLM instructions embedded in HTML. These are, in essence, accessibility accommodations. But they're being offered cheerfully to AI systems after decades of being denied to blind and low-vision users. From lane markings repainted for autonomous vehicles to sidewalk robots competing with wheelchair users for the curb cuts activists once poured by hand, a pattern is emerging in which infrastructure that disabled people fought for materializes only when a well-capitalized technology needs the same thing. In this talk, I trace this pattern across physical and digital infrastructure and examine why machine-readability is not the same as accessibility. This distinction reflects differences between how machines ingest content and how people perceive, navigate, and comprehend it. As AI systems become a primary audience for educational content and infrastructure, I invite the audience to consider what these changes suggest for learning environments: whose ways of reading, knowing, and accessing are designed for, and whose are treated as optional.

Speaker biography:

Jonathan Zong is an Assistant Professor of Information Science at the University of Colorado Boulder, where he leads the Data & Design Group. He is also a Faculty Associate at Harvard's Berkman Klein Center for Internet and Society. In his research, Jonathan partners with blind collaborators and study participants to co-design accessible interfaces for non-visual data exploration. He has also developed systems and frameworks for technology refusal and data ethics, supporting communities to resist harmful data collection practices. His work has been recognized by the MIT Morningside Academy for Design Fellowship, the Paul and Daisy Soros Fellowship for New Americans, and Forbes 30 Under 30 Scientists.

Embodied AI: Designing Inclusive Digital Ecologies

Ronnie Videla¹, Dor Abrahamson¹ and Marcelo Chávez²

¹University of California Berkeley, USA.

²Universidad de La Serena, FABLAB, Chile.

Abstract. Contemporary approaches to Artificial Intelligence in Education (AIED) are predominantly grounded in representational models that emphasize prediction, personalization, and learning optimization. This paper proposes an alternative perspective by integrating Embodied Interaction, Structural Coupling, Ecological Dynamics, Embodied Design, and the Special Education Embodied Design (SpEED) framework. We introduce Embodied AI as an ecological approach to Human–Computer Interaction in which AI is understood not as a source of instruction or knowledge delivery, but as a participant in embodied systems of practice. Within this perspective, meaning emerges through sociomaterial coordination among learners, technologies, environments, and cultural practices. The framework is illustrated through two inclusive mathematics learning environments, QUAD and Graph-Math, originally designed for blind learners. Rather than treating accessibility as accommodation, these cases demonstrate how inclusive design can function as a source of epistemic innovation, expanding opportunities for participation and revealing alternative pathways for mathematical meaning-making. We argue that the future challenge of educational AI is not prediction but participation, and that Embodied AI provides a foundation for designing inclusive ecological systems of Human–Computer Interaction in which intelligence emerges through evolving relations among bodies, technologies, and environments.

1 INTRODUCTION

Artificial intelligence is rapidly reshaping contemporary educational systems. Intelligent tutoring systems, adaptive learning platforms, learning analytics, and generative AI applications are typically grounded in the assumption that learning can be enhanced through increasingly accurate models of learner behavior and the efficient delivery of instructional support (OECD, 2025). While these approaches have produced powerful technological innovations, they remain largely informed by representational and information-processing theories of cognition, in which intelligence is understood as the manipulation of internal models and learning as the acquisition of abstract knowledge structures. In contrast, contemporary developments in embodied cognition, ecological psychology, enactivism, and Human–Computer Interaction suggest that cognition does not reside exclusively within individuals or computational systems, but emerges through ongoing interactions among bodies, environments, artifacts, and sociocultural practices (Varela, Thompson, & Rosch, 1991; Thompson, 2007; Gibson, 1979; Chemero, 2009). From this perspective, meaning arises not through the passive reception of

information but through active participation in systems of perception and action.

These ideas have been further developed through theories of structural coupling, ecological dynamics, and embodied learning. Structural coupling proposes that cognition emerges through recurrent histories of organism–environment interaction, where learning is understood as the transformation of relational dynamics rather than the internalization of representations (Maturana & Varela, 1987; Varela, 1997). Ecological dynamics similarly conceptualizes learning as an emergent process arising from interactions among agent, task, and environmental constraints (Araújo & Davids, 2011; Newell & Ranganathan, 2010). Within mathematics education, these commitments have inspired Embodied Design, a pedagogical framework that investigates how mathematical concepts emerge through coordinated interactions among bodies, artifacts, movements, and symbolic practices (Abrahamson & Sánchez-García, 2016). More recently, the Special Education Embodied Design (SpEED) framework has extended these insights into inclusive educational contexts by foregrounding the embodied resources of learners whose sensory, cognitive, or communicative experiences are often marginalized within conventional learning environments (Tancredi et al., 2021, 2022; Abrahamson et al., 2023).

Despite these advances, their implications for artificial intelligence and Human–Computer Interaction remain underdeveloped (Videla, Penny & Ross, 2025). Educational AI continues to be predominantly framed in terms of prediction, personalization, and automated instruction, while HCI often focuses on usability, efficiency, or information exchange. Such perspectives overlook the possibility that intelligent technologies may participate directly in the emergence of meaning through embodied systems of practice. This paper addresses this gap by proposing the concept of Embodied AI. Drawing upon Embodied Interaction (Dourish, 2001), Structural Coupling, Ecological Dynamics, Embodied Design, and SpEED, we conceptualize AI not as a source of knowledge delivery but as an ecological participant within sociomaterial learning systems. From this perspective, AI contributes to the ongoing reorganization of affordances, constraints, modalities, and opportunities for action through which meaning emerges. We illustrate the explanatory potential of this framework through two inclusive mathematics learning environments, QUAD and Graph-Math, and argue that the future challenge of educational AI is not prediction but participation: designing intelligent technologies that support embodied, inclusive, and meaningful forms of Human–Computer Interaction.

2 FROM REPRESENTACION TO PARTICIPATION

Contemporary developments in Artificial Intelligence for Education (AIED) have been largely driven by advances in machine learning, predictive analytics, adaptive systems, and generative models. Intelligent tutoring systems, learner models, recommendation engines, and generative AI applications are typically designed around the capacity to infer internal states and deliver personalized instructional interventions. Within this paradigm, cognition is often understood as the processing and manipulation of information, while learning is conceived as the acquisition or refinement of internal representations.

Over the past several decades, a growing body of research has challenged the adequacy of representational accounts for understanding human cognition and learning (Gallagher, 2017). Perspectives emerging from embodied cognition, ecological psychology, enactivism, and phenomenology have argued that cognition cannot be reduced to internal information processing (Hutto y Myin, 2013). A foundational contribution to this shift was provided by ecological psychology. Gibson (1979) rejected the notion that perception requires the internal reconstruction of an external world through mental representations. Instead, he proposed that organisms directly perceive affordances, opportunities for action available within specific environmental conditions (Turvey, 1992). Perception is therefore inseparable from action. Organisms do not passively receive information and subsequently interpret it, they actively explore environments in ways that reveal possibilities for engagement (Malafouris, 2013). Meaning emerges not from internal representations but from practical relations between organisms and their surroundings.

Similar critiques emerged within embodied and enactive approaches to cognition. Varela, Thompson, and Rosch (1991) argued that cognition should be understood as embodied action rather than symbolic computation. From an enactive perspective, organisms do not represent a pre-given world, they enact meaningful worlds through histories of structural coupling with their environments (Maturana y Varela, 1980; Hutto y Myin, 2017). These developments have profound implications for Human–Computer Interaction. If cognition is fundamentally embodied, situated, and relational, then interaction cannot be understood merely as information exchange between users and computational systems. Instead, interaction becomes a process through which meaning emerges in practice (Dourish, 2001). This shift is particularly significant in educational contexts, where learning technologies are often evaluated according to their capacity to deliver information, provide feedback, or improve measurable outcomes. Such criteria may obscure more fundamental questions concerning how technologies participate in the constitution of activity itself.

Learners do not encounter educational environments as isolated cognitive agents. They participate in complex ecologies involving bodies, artifacts, representations, cultural practices, and other people. Understanding these dynamics requires moving beyond representational accounts of cognition toward a participation-centered understanding of interaction. In this paper, we adopt participation rather than representation as the primary unit of analysis. Participation refers to the ways in which

learners become engaged in embodied systems of practice through ongoing interactions with people, technologies, materials, and environments. From this perspective, the central challenge for educational technology is not how to improve the transmission of information, but how to support meaningful forms of participation. This shift provides the conceptual foundation for rethinking artificial intelligence as a component of embodied learning ecologies rather than as an external source of instruction or knowledge. This reconceptualization has important implications for understanding interaction, learning, and technology. If cognition is fundamentally embodied and relational, interaction cannot be reduced to the transmission of information between independent entities. Human activity unfolds within ecological systems composed of bodies, artifacts, representations, technologies, cultural practices, and other people, all of which participate in shaping possibilities for action.

3 EMBODIED INTERACTION AND HUMAN-COMPUTER INTERACTION

The shift from representation to participation requires a corresponding reconceptualization of Human–Computer Interaction. If cognition is understood as emerging through embodied engagement with the world, then interaction cannot be reduced to the exchange of information between users and computational systems. Instead, interaction becomes a process through which meaning is enacted, negotiated, and transformed within ongoing systems of practice. A particularly influential contribution to this perspective was provided by Dourish (2001), whose theory of embodied interaction challenged dominant computational models of HCI by grounding interaction in phenomenological accounts of human experience.

Drawing upon Heidegger, Merleau-Ponty, Schutz, and Wittgenstein, Dourish argued that interaction should not be understood primarily in terms of representations, commands, or information transfer. Rather, interaction emerges through embodied participation in meaningful activities. Technologies become intelligible not because users possess internal models of how they operate, but because they become incorporated into practical engagements with the world. Meaning is not located inside computational representations or mental structures; it emerges through situated action.

Central to Dourish's account is the concept of coupling. Human beings do not simply use tools; they become coupled with them through practice. As tools become integrated into activity, they cease to function as external objects and instead become components of ongoing action. Dourish illustrates this point through the example of writing with a pen. Skilled writers do not consciously coordinate individual finger movements or monitor the pen as an object, rather, they become engaged in the activity of writing itself. Interaction, therefore, is not a relationship between separate entities but a process of mutual coordination among bodies, artifacts, and environments.

This understanding challenges persistent assumptions within contemporary digital education and artificial intelligence. Many educational technologies continue to be designed as systems that deliver information, provide feedback, or optimize user

performance. Such approaches implicitly position technology as an external mediator between knowledge and learner. Dourish's (2001) perspective argues that meaning is not imposed by designers, encoded within interfaces, or determined by computational representations, therefore, interaction cannot be fully specified in advance. Technologies create possibilities for action, but the significance of those possibilities emerges through participation in practice. This perspective resonates strongly with ecological approaches to perception and action, where affordances are understood not as fixed properties of environments but as relational opportunities that emerge through organism–environment engagement.

The educational significance of a technology cannot be understood solely in terms of its informational content or instructional effectiveness. It must also be understood in terms of how it reorganizes opportunities for action, perception, exploration, and participation. Embodied interaction also challenges the notion of context as an external container surrounding activity. For Dourish (2001), context and activity are mutually constitutive. This insight has important implications for educational technology design, since learning environments cannot be understood as neutral settings within which learning occurs. Instead, they are dynamically enacted through ongoing interactions among learners, teachers, technologies, materials, and institutional practices.

Taken together, these ideas suggest a fundamentally different understanding of Human–Computer Interaction. Interaction is not information exchange. It is embodied participation in systems of practice. Technologies are not neutral tools that mediate between users and information. Rather than conceptualizing AI as an autonomous source of knowledge, prediction, or instruction, we propose that artificial intelligence should be understood as part of the sociomaterial ecology through which interaction and meaning emerge. This proposal provides the foundation for integrating embodied interaction with ecological dynamics, structural coupling, Embodied Design, and inclusive educational frameworks.

4 STRUCTURAL COUPLING AND RELATIONAL DYNAMICS OF INTERACTION

A complementary account of interaction is provided by the Biology of Cognition developed by Maturana and Varela (1980, 1987). This framework begins with a radical departure from representational theories of mind by grounding cognition in the organizational properties of living systems. A central implication of this view is that living systems are structurally determined systems. At any given moment, all changes that occur within an organism are determined by its current structure rather than by external inputs. Environmental events do not instruct organisms, transmit information into them, or determine their internal states. Rather, environmental perturbations trigger structural changes whose specific consequences depend on the organism's own organization and structural history. Consequently, cognition cannot be understood as the reception or processing of information from an objective external world. What an organism

does in response to a perturbation is always determined by its own structure at that moment.

Within this framework, structural coupling describes the historical process through which organism and environment undergo mutually congruent transformations through recurrent interactions (Maturana y Varela, 1980). Through continued engagement, organisms become increasingly attuned to the regularities of their environments, while environments simultaneously become structured through the activities of organisms. Over time, these reciprocal adjustments generate domains of viable interaction in which organism and environment become dynamically coordinated. Structural coupling is therefore not a connection between pre-existing entities but an ongoing process through which relational coherence emerges and is maintained.

From this perspective, cognition is not the representation of an external reality but the enactment of a meaningful world through histories of interaction. Organisms do not discover a pre-given world; they bring forth domains of significance through recurrent cycles of perception and action. Meaning emerges within the relational dynamics of coupling itself, cognition is therefore fundamentally embodied, situated, and historical. What an organism can perceive, discriminate, and coordinate depends on the history of couplings through which its current structure has been constituted. These ideas have been further developed within contemporary enactivist and ecological approaches to cognition. Radical and extensive enactivism argue that cognition should not be understood as the manipulation of internal representational content but as the unfolding of embodied organism–environment dynamics (Hutto & Myin, 2013; Hutto et al., 2014). Cognitive processes are not internal events occasionally extended into the world; rather, they are constitutively relational processes distributed across evolving systems of interaction. This shift further reinforces the view that cognition is enacted through participation rather than generated inside individual minds.

Recent developments in Embodied AI have extended this framework by distinguishing among constitutive, enabling, and contextual factors involved in cognitive phenomena (Videla, Penny and Ross, 2025). Constitutive factors are those processes that directly participate in the production of a phenomenon. Enabling factors create the conditions that make the phenomenon possible, while contextual factors modulate its expression without directly producing it. Learning, from this perspective, emerges through dynamic interactions among these domains rather than from isolated causal mechanisms. This understanding has significant implications for Human–Computer Interaction. Technologies should not be viewed as external devices that transmit information into learners. Nor should artificial intelligence be understood as a system that directly causes learning. Instead, technological systems participate as components within broader ecologies of structural coupling. They introduce perturbations, reorganize constraints, and modify the affordances available within ongoing activity. Their influence is always mediated by the structural dynamics of the systems in which they participate.

Consequently, the educational significance of artificial intelligence lies not in its capacity to represent knowledge more accurately or provide more efficient feedback. Rather, its significance lies in its capacity to participate in evolving systems of structural coupling that support new forms of embodied interaction. AI becomes part of the relational dynamics through which learners, teachers, artifacts, and environments continuously co-constitute one another. Understanding these dynamics requires moving beyond representational accounts of technology toward an ecological conception of interaction grounded in participation, coupling, and the emergence of meaning.

5 EMBODIED DESIGN AND INCLUSIVE LEARNING ECOLOGIES

Embodied Design emerged as a response to the limitations of representationalist theories of mathematical learning. Rather than conceiving learning as the progressive acquisition of abstract representations, this framework proposes that mathematical understanding arises through the reorganization of perception–action systems within carefully structured environments (Abrahamson, 2009, 2014). Drawing on embodied cognition, ecological psychology, enactivism, and dynamical systems theory, Embodied Design conceptualizes learning as a relational, situated, and embodied process grounded in interactions among learners, artifacts, environments, and cultural practices.

Within this perspective, mathematical concepts emerge through the stabilization of perceptual and motor coordinations that become meaningful within activity (Abrahamson & Sánchez-García, 2016). Mathematical knowing is therefore not primarily a matter of acquiring symbolic content but of developing new ways of perceiving, acting, discriminating, and coordinating within mathematically organized environments. This view aligns closely with extensive enactivism, which rejects the assumption that cognition is fundamentally representational and instead understands cognition as world-involving from the outset (Hutto & Myin, 2013; Hutto, 2017).

Rather than treating mathematical ideas as external abstractions, Embodied Design creates conditions in which learners can experience conceptual relations through embodied activity. Learning emerges through a gradual alignment between evolving sensorimotor coordinations and culturally valued forms of mathematical reasoning. Embodied Design operationalizes these commitments through the deliberate configuration of learning ecologies. Drawing on ecological dynamics, learning is understood as emerging from interactions among agent constraints, environmental constraints, and task constraints (Newell, 1986; Newell & Ranganathan, 2010). Agent constraints include learners’ embodied histories and perceptual capacities; environmental constraints involve material, digital, spatial, and social conditions; and task constraints refer to goals, rules, and forms of feedback. Learning is not produced through information transmission but through adaptive reorganization within these interacting constraints.

Within this framework, perception is understood as attunement to affordances (Gibson, 1979). Educational design therefore does not impose meaning but structures environments in which

meaningful affordances can emerge. Designers create opportunities for learners to encounter and stabilize mathematically relevant relations through embodied engagement. In this sense, Embodied Design can be understood as the pedagogical operationalization of ecological dynamics, translating principles of perception–action coupling into educational practice (Abrahamson et al., 2023).

A central contribution of this framework is the concept of attentional anchors (Abrahamson et al., 2011; Abrahamson & Sánchez-García, 2016). Attentional anchors are perceptually salient and action-relevant structures that learners spontaneously develop in order to coordinate activity within mathematically organized tasks. Rather than functioning as symbolic representations, they serve as ecological control structures that stabilize perception–action loops and support the detection of mathematically relevant invariants. Through interaction with materials, technologies, and social partners, attentional anchors become progressively refined, enabling increasingly sophisticated forms of coordination that later ground formal mathematical reasoning (Abrahamson & Bakker, 2016; Abrahamson et al., 2020).

This perspective has important implications for educational technology. Technologies are not merely vehicles for delivering information or representing concepts. Instead, they participate directly in the organization of perception–action systems and become components of evolving ecologies of structural coupling. Their educational significance lies not in the content they communicate but in the forms of participation they make possible. From this standpoint, artificial intelligence can be reconceptualized as a dynamic modulator of ecological constraints. Rather than functioning as a source of explanations or instructional authority, AI can reorganize affordances, amplify perceptual cues, support multimodal engagement, and sustain emerging attentional anchors without eliminating exploratory activity. AI thus becomes part of a distributed cognitive ecology that supports conceptual emergence while preserving learner agency. Its role is not to replace human reasoning but to participate in the ongoing reconfiguration of conditions under which meaningful action becomes possible.

The relevance of this perspective becomes particularly evident in inclusive educational contexts. Conventional educational environments often privilege narrow sensory, motor, and representational pathways, marginalizing learners whose embodied experiences diverge from normative assumptions. Embodied Design challenges this orientation by treating bodily diversity not as a limitation but as a resource for design. Different sensory and cognitive profiles reveal alternative affordances, alternative coordinations, and alternative pathways toward conceptual understanding. In this sense, inclusion is not merely a matter of accessibility but a source of epistemic innovation. Research with blind and visually impaired learners has demonstrated that accessibility can generate new theoretical insights into perception, coordination, and learning (Abrahamson et al., 2019). More recently, Abrahamson et al. (2023) has described this orientation as Inclusive Design for Expansive Theory, arguing that educational researchers can develop more powerful theories of learning by designing from the embodied resources of diverse learners. Rather than adapting existing

models to accommodate difference, this perspective treats diversity itself as a generative condition for theoretical and pedagogical innovation. Viewed through this lens, Embodied Design is not merely a pedagogical methodology for mathematics education. It is an ecological design framework for intentionally configuring sociomaterial conditions under which new forms of perception, participation, structural coupling, and meaning-making can emerge. This orientation provides the conceptual bridge toward the Special Education Embodied Design (SpEED) framework and, ultimately, toward our conception of Embodied AI as an ecological approach to inclusive Human–Computer Interaction.

6 THE SPEED FRAMEWORK: TOWARD AN ECOLOGICAL MODEL OF INCLUSIVE HUMAN-COMPUTER INTERACTION

While Embodied Design provides a powerful framework for understanding how mathematical meanings emerge through embodied participation, it offers fewer explicit resources for addressing the diversity of sensory, motor, and cognitive experiences present in contemporary educational environments. This challenge becomes particularly important when designing technologies intended to support learners whose perceptual and interactional profiles diverge from normative assumptions. In response to this challenge, researchers have developed the Special Education Embodied Design (SpEED) framework, an extension of Embodied Design that places embodied diversity at the center of educational innovation (Tancredi et al., 2021, 2022).

SpEED emerged from a growing recognition that conventional approaches to inclusion often treat accessibility as a process of adapting existing learning environments to accommodate individual differences. In contrast, SpEED begins from the assumption that perceptual diversity constitutes a productive resource for educational design. Learners who are blind, autistic, neurodivergent, or otherwise marginalized frequently reveal dimensions of interaction that remain invisible within standard educational settings. Rather than viewing these learners through a deficit lens, SpEED treats their embodied experiences as generative starting points for the design of new learning ecologies. Abrahamson et al. (2019) demonstrated how tactile mathematical activities involving blind learners challenge traditional assumptions about mathematical perception and reveal alternative pathways toward conceptual understanding.

Subsequent work argued that accessibility should not be understood merely as accommodation but as a source of theoretical and pedagogical innovation (Tancredi et al., 2021, 2022; Abrahamson, 2023). This shift reframes inclusion as a process of expanding the ecological possibilities for participation rather than adapting learners to existing instructional structures. SpEED operationalizes this perspective through three interrelated design parameters: media, modalities, and semiotic modes. Together, these dimensions provide a framework for analyzing and designing embodied learning environments while also offering a powerful model for understanding Human–Computer Interaction within inclusive educational ecologies.

From Tancredi et al (2022) the first parameter, media, refers to the material and technological resources that participate in activity. Media include physical artifacts, digital interfaces, tangible manipulatives, sonification systems, sensors, intelligent technologies, and other components that become integrated into embodied systems of practice. Importantly, media are not understood as passive instructional supports. Consistent with Dourish’s (2001) conception of embodied interaction, they participate directly in the constitution of activity itself. Media shape what can be perceived, coordinated, explored, and enacted within a learning ecology.

The second parameter, modalities, refers to the sensory and motor channels through which learners engage with the environment. Vision, touch, audition, proprioception, kinesthetic awareness, gesture, and movement all constitute modalities through which perception and action become organized. SpEED recognizes that learners do not access educational environments through identical sensory pathways. Different modalities reveal different affordances and support different forms of participation. Consequently, inclusive design requires attention not only to what learners know but also to how they perceive, act, and coordinate within educational activity.

The third parameter, semiotic modes, concerns the representational and communicative forms through which meaning is expressed and negotiated (Vygotsky, 1962). These include gestures, embodied actions, diagrams, spoken language, tactile representations, symbolic notations, and digital visualizations. Within SpEED, semiotic modes are not viewed as endpoints of cognition but as resources that emerge from and remain grounded in embodied activity. Mathematical symbols, for example, are understood as culturally stabilized forms of coordination that become meaningful through their relation to embodied action and perception (Abrahamson, 2021). From the perspective developed in this paper, SpEED can be understood as an ecological model of Human–Computer Interaction. Its significance extends beyond special education because it provides a concrete vocabulary for analyzing how technologies participate in embodied systems of practice. This ecological interpretation of SpEED also provides a bridge between inclusive design and Human–Computer Interaction. Inclusion is no longer understood as a specialized concern applicable only to particular learner populations. Instead, it becomes a methodological strategy for expanding the range of affordances available within sociomaterial systems. By designing from the embodied resources of diverse learners, researchers gain access to richer forms of interaction, more flexible configurations of participation, and ultimately more expansive theories of cognition and learning.

7 EMBODIED AI: AN ECOLOGICAL FRAMEWORK FOR INCLUSIVE HUMAN-COMPUTER INTERACTION

The theoretical perspectives reviewed thus far converge on a common proposition: cognition does not reside exclusively within individual minds or computational systems but emerges through embodied participation in sociomaterial ecologies. From perspectives Embodied Interaction (Dourish, 2001), Ecological

Dynamics (Araújo & Davids, 2011), Structural Coupling (Maturana & Varela, 1987). Embodied Design (Abrahamson, 2014) and SpEED (Tancredi et al., 2022)., we propose Embodied AI as an ecological framework for Human–Computer Interaction. Rather than conceptualizing artificial intelligence as a source of instruction, prediction, personalization, or knowledge delivery, Embodied AI understands AI systems as participants within embodied systems of practice. Their role is not to replace human cognition, optimize learning, or automate decision making. Instead, intelligent systems participate in the ongoing reorganization of affordances, constraints, modalities, and semiotic resources through which meaning emerges. Consequently, the central question shifts from “What knowledge does AI provide?” to “How does AI participate in the organization of meaningful activity?”

This framework becomes particularly relevant within inclusive educational contexts, where diversity in sensory, motor, and cognitive experience reveals alternative pathways for participation and meaning-making. To illustrate the explanatory potential of Embodied AI, we examine two embodied mathematics learning environments: QUAD and Graph-Math. These systems were originally developed to support blind learners, yet their significance extends beyond accessibility. They demonstrate how designing from marginalized embodied experiences can generate more expansive forms of Human–Computer Interaction for all learners.

8 ILLUSTRATIVE CASES OF EMBODIED AI

QUAD: Reconfiguring Geometric Interaction Through Embodied AI



Figure 1: QUAD Case

QUAD was initially conceived as an embodied mathematics environment that enables blind and low-vision learners to explore geometric invariants through tactile and proprioceptive interaction. Unlike conventional approaches that translate visual diagrams into verbal descriptions, QUAD reorganizes the learning ecology itself. Geometry is no longer treated as a visual object to be represented but as a set of relational structures that can be enacted through embodied exploration. The system consists of a sensorized quadrilateral whose sides and vertices can be manipulated physically. These movements are continuously processed through a mechatronic architecture that

couples tactile action with computational interpretation and auditory feedback. Within a traditional Human–Computer Interaction framework, the device might be understood as an interface for transmitting geometric information. From an Embodied AI perspective, however, QUAD functions as a sociomaterial ecology in which learners, artifacts, computational processes, and mathematical relations become structurally coupled through activity.

Analyzed through SpEED, QUAD reveals how embodied participation emerges through the interaction of multiple design parameters. As media, the manipulable artifact, sensors, computational infrastructure, and auditory feedback system operate as active components of the cognitive ecology. As modalities, tactile, proprioceptive, kinesthetic, and auditory channels become primary resources for perception and coordination. As semiotic modes, geometric categories such as square, rectangle, trapezoid, or parallelogram emerge only after learners stabilize embodied relations among movement, shape transformation, and auditory cues. Within this ecology, AI does not function as an evaluator of correctness. Rather, it participates in the stabilization of interaction. By identifying relational transformations and maintaining coherent correspondences between tactile manipulations and auditory responses, the intelligent system helps sustain the perception–action dynamics through which geometric invariants become available for exploration. The significance of AI lies not in providing answers but in supporting the ecological conditions that allow learners to discover and enact mathematical relations.

Importantly, QUAD illustrates a broader principle of expansive inclusive design. What began as a design for blind learners reveals fundamental aspects of geometric cognition that conventional visual instruction often obscures. The system demonstrates that geometric concepts are not fundamentally visual representations but embodied invariants emerging through interaction. Designing from the perceptual resources of blind learners therefore expands the theoretical understanding of geometry itself, generating richer possibilities for participation across diverse learner populations.

Graph-Math: Embodied Functional Reasoning Through Multimodal Coordination



Figure 2: Graph Math Case

Graph-Math extends these principles into the domain of analytic geometry and functional reasoning. Originally developed to support blind students in exploring Cartesian relations without

reliance on visual graphs, the system establishes a dynamic correspondence between bodily action, tactile exploration, computational processing, and mathematical conceptualization. At the center of the environment is a 13×13 sensorized tactile matrix composed of 169 pressure-sensitive nodes connected to an Arduino-based computational architecture. Learners physically position markers within the grid, generating real-time mappings onto a digital Cartesian plane. Through repeated interaction, coordinate relations, slopes, intercepts, and functional transformations become available as embodied experiences rather than abstract visual objects.

Within the Embodied AI framework, Graph-Math is not a tool for translating visual graphs into alternative sensory formats. Instead, it constitutes an ecological environment in which mathematical meanings emerge through the coordination of action, perception, and representation. The learner contributes embodied histories, exploratory strategies, and sensorimotor capacities. The task involves investigating relationships among coordinates, rates of change, and functional trajectories. The environment consists of a tightly coupled network of tactile surfaces, computational processing, auditory feedback, and graphical representation.

The SpEED framework illuminates these dynamics. As media, Graph-Math integrates tactile hardware, computational infrastructure, sonification systems, and graphical displays. As modalities, tactile, kinesthetic, auditory, and visual channels become coordinated within a shared perception-action system. As semiotic modes, symbolic notation, coordinate pairs, graphs, gestures, and spoken language emerge as interconnected resources for mathematical reasoning. The role of AI within Graph-Math differs significantly from that of conventional intelligent tutoring systems. Rather than diagnosing misconceptions or delivering corrective instruction, the AI maintains structural correspondences across modalities and representations. When learners position markers inconsistently with the targeted mathematical relations, the system generates exploratory prompts that invite further investigation rather than prescribing solutions. AI thus functions as an ecological participant that helps sustain coherent relations among embodied actions, tactile configurations, auditory cues, and graphical representations.

Through repeated interaction, learners develop attentional anchors that enable them to perceive slope as inclination, intercept as boundary relation, and functional change as a dynamic transformation of trajectories. Mathematical meaning emerges not from the interpretation of representations but from embodied participation in a multimodal ecology of action. As in QUAD, the significance of the design extends beyond accessibility. By foregrounding tactile and kinesthetic interaction, Graph-Math reveals alternative epistemic pathways that enrich mathematical learning for all learners. Embodied AI as Inclusive Human-Computer Interaction Together, QUAD and Graph-Math illustrate a shift from instructional AI toward ecological AI. In both cases, artificial intelligence does not serve as an evaluator, or predictive system. Instead, it participates in embodied ecologies of interaction, supporting the emergence of affordances, stabilizing multimodal coordinations, and sustaining opportunities for participation.

7 CONCLUSIONS

This paper has proposed Embodied AI as an ecological framework for inclusive Human-Computer Interaction. Drawing upon Embodied Interaction, Structural Coupling, Ecological Dynamics, Embodied Design, and the SpEED framework, we have argued that artificial intelligence should not be understood primarily as a mechanism for prediction, personalization, or automated instruction. Instead, AI can be conceived as a participant within embodied systems of practice that contribute to the ongoing reorganization of affordances, constraints, modalities, and opportunities for action through which meaning emerges. The cases of QUAD and Graph-Math illustrate how intelligent technologies can support multimodal participation and embodied forms of mathematical reasoning while preserving learner agency. More broadly, they demonstrate that designing from the embodied resources of marginalized learners does not merely increase accessibility; it expands the theoretical understanding of cognition, interaction, and learning itself. Ultimately, the contribution of Embodied AI lies in shifting the focus of educational technology from optimization to participation, and from information delivery to the creation of inclusive sociomaterial ecologies in which learners, technologies, and environments jointly enact meaningful worlds.

7 DISCUSSIONS: WHAT MAKES AI EMBODIED?

A central question emerging from this work concerns why the proposed framework should be described as Embodied AI. After all, both QUAD and Graph-Math could function without sophisticated artificial intelligence. Their educational value does not depend on automated instruction, predictive analytics, or learner modeling. This observation reveals an important conceptual distinction. In conventional educational technology, AI is typically defined by its capacity to classify, predict, recommend, or personalize. Within the Embodied AI framework, however, AI is defined by the role it plays within embodied systems of practice.

Artificial intelligence becomes embodied not when it simulates human cognition, but when it participates in the organization of perception-action systems. More specifically, AI becomes embodied when it sustains multimodal correspondences across tactile, auditory, visual, and kinesthetic channels; supports the emergence and stabilization of attentional anchors; modulates ecological constraints without eliminating exploratory activity; reorganizes affordances that shape possibilities for participation; and participates in evolving histories of structural coupling among learners, artifacts, and environments. In this sense, AI is not conceived as an external tutor or intelligent authority but as an ecological participant within sociomaterial systems of interaction.

This perspective extends Dourish's notion of embodied interaction beyond traditional Human-Computer Interaction by incorporating contemporary developments in ecological dynamics, enactivism, and inclusive design. Rather than evaluating AI according to the accuracy of its predictions or the efficiency of its instructional interventions, Embodied AI redirects attention toward the ways intelligent technologies participate in the emergence of meaningful activity. The significance of AI therefore lies not in what it knows, but in how

it contributes to the ongoing organization of participation, coordination, and sense-making. Such a shift opens new directions for Human–Computer Interaction research, particularly within collaborative, creative, and inclusive learning environments where meaning emerges through distributed systems of embodied action rather than through information transmission alone.

REFERENCES

- [1] OECD. *The Impact of Digital Technologies on Students' Learning and Well-Being*. OECD Publishing, Paris, 2025.
- [2] F. J. Varela, E. Thompson and E. Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, Cambridge, MA, 1991.
- [3] E. Thompson. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press, Cambridge, MA, 2007.
- [4] J. J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.
- [5] A. Chemero. *Radical Embodied Cognitive Science*. MIT Press, Cambridge, MA, 2009.
- [6] H. R. Maturana and F. J. Varela. *The Tree of Knowledge: The Biological Roots of Human Understanding*. Shambhala, Boston, 1987.
- [7] F. J. Varela. "The Naturalization of Phenomenology as the Transcendence of Nature." *Alter*, 5, pp. 355–385, 1997.
- [8] D. Araújo and K. Davids. "What Exactly Is Acquired During Skill Acquisition?" *Journal of Consciousness Studies*, 18(3–4), pp. 7–23, 2011.
- [9] K. M. Newell and R. Ranganathan. "Instructions as Constraints in Motor Skill Acquisition." In: *Motor Learning in Practice*. Routledge, London, pp. 17–32, 2010.
- [10] D. Abrahamson and R. Sánchez-García. "Learning Is Moving in New Ways: The Ecological Dynamics of Mathematics Education." *Journal of the Learning Sciences*, 25(2), pp. 203–239, 2016. <https://doi.org/10.1080/10508406.2015.1125287>
- [11] S. Tancredi, R. S. Y. Chen, C. M. Krause, D. Abrahamson and P. Gomez. "Getting Up to SpEED: Special Education Embodied Design for Sensorially Equitable Inclusion." *Education Sciences & Society*, 12(1), pp. 114–136, 2021. <https://doi.org/10.3280/ess1-2021oa11818>
- [12] S. Tancredi, R. S. Y. Chen, C. M. Krause and Y.-T. Siu. "The Need for SpEED: Reimagining Accessibility Through Special Education Embodied Design." In S. L. Macrine and J. M. B. Fugate (Eds.), *Movement Matters: How Embodied Cognition Informs Teaching and Learning*, pp. 197–216. MIT Press, Cambridge, MA, 2022.
- [13] D. Abrahamson, S. Tancredi, R. S. Y. Chen, V. J. Flood and E. Dutton. "Embodied Design of Digital Resources for Mathematics Education: Theory, Methodology, and Framework of a Pedagogical Research Program." In B. Pepin, G. Gueudet and J. Choppin (Eds.), *Handbook of Digital Resources in Mathematics Education*. Springer, Cham, 2023. https://doi.org/10.1007/978-3-030-95060-6_8-1
- [14] R. Videla, S. Penny and W. Ross. "If You Can't Dance Your Program, You Can't Write It": Challenges and Implications for AI in Education. *ACM Transactions on Computing Education*, 26(3), pp. 1–12, 2025. <https://doi.org/10.1145/3759257>
- [15] P. Dourish. *Where the Action Is: The Foundations of Embodied Interaction*. MIT Press, Cambridge, MA, 2001.
- [16] S. Gallagher. *Enactivist Interventions: Rethinking the Mind*. Oxford University Press, Oxford, 2017.
- [17] D. D. Hutto and E. Myin. *Radicalizing Enactivism: Basic Minds Without Content*. MIT Press, Cambridge, MA, 2013.
- [18] M. Turvey. "Affordances and Prospective Control: An Outline of the Ontology." *Ecological Psychology*, 4(3), pp. 173–187, 1992.
- [19] L. Malafouris. *How Things Shape the Mind: A Theory of Material Engagement*. MIT Press, Cambridge, MA, 2013.
- [20] H. R. Maturana and F. J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company, Dordrecht, 1980.
- [21] D. D. Hutto. "Getting into Predictive Processing's Great Guessing Game." *Synthese*, 195(6), pp. 2647–2666, 2018. <https://doi.org/10.1007/s11229-017-1385-0>
- [22] D. D. Hutto, M. D. Kirchhoff and E. Myin. "Extensive Enactivism." *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pp. 676–681, 2014.
- [23] D. Abrahamson. "Embodied Design: Constructing Means for Constructing Meaning." *Educational Studies in Mathematics*, 70(1), pp. 27–47, 2009. <https://doi.org/10.1007/s10649-008-9137-1>
- [24] D. Abrahamson. "Building Educational Activities for Understanding: An Elaboration on the Embodied Design Framework and Its Epistemic Grounds." *International Journal of Child–Computer Interaction*, 2(1), pp. 1–16, 2014. <https://doi.org/10.1016/j.ijcci.2014.07.002>
- [25] D. Abrahamson, D. Trninic, J. F. Gutiérrez, K. Huthmacher and R. G. Lee. "Hooks and Shifts: A Dialectical Study of Mediated Discovery in Embodied Mathematics Learning." *Technology, Knowledge and Learning*, 16(1), pp. 55–85, 2011. <https://doi.org/10.1007/s10758-011-9177-0>
- [26] D. Abrahamson and A. Bakker. "Making Sense of Movement in Embodied Design for Mathematics Learning." *Cognitive Research: Principles and Implications*, 1(33), 2016. <https://doi.org/10.1186/s41235-016-0034-3>
- [27] D. Abrahamson, M. J. Nathan, C. Williams-Pierce, C. Walkington, E. R. Ottmar, H. Soto and M. W. Alibali. "The Future of Embodied Design for Mathematics Teaching and Learning." *Frontiers in Education*, 5, 147, 2020. <https://doi.org/10.3389/feduc.2020.00147>
- [28] D. Abrahamson, V. J. Flood, J. A. Miele and Y.-T. Siu. "Enactivism and Ethnomethodological Conversation Analysis as Tools for Expanding Universal Design for Learning: The Case of Visually Impaired Mathematics Students." *ZDM Mathematics Education*, 51(2), pp. 291–303, 2019. <https://doi.org/10.1007/s11858-018-0998-1>
- [29] D. Abrahamson. "Grasp Actually: An Evolutionist Argument for Enactivist Mathematics Education." *Human Development*, 65(1), pp. 1–17, 2021. <https://doi.org/10.1159/000515680>
- [30] L. S. Vygotsky. *Thought and Language*. MIT Press, Cambridge, MA, 1962.

Can an LLM build an application to automatically assess students' interpretations of diagrams?

Grecia Garcia Garcia¹ and Peter C-H. Cheng¹

Abstract. RIST is a theory of the mental structures of external representations, and RISN is its notation [1, 5]. In this exploratory study, we developed a pilot web application with the assistance of Claude's Sonnet 4.6 to: (a) investigate whether RIST/N can be used for pedagogic evaluation of students' diagrammatic interpretations, and (b) identify the challenges that must be addressed to make such a system effective for educators. This application processes student interpretations of diagrams and provides feedback to the teacher on where students are experiencing difficulties. The system has three inputs: (i) a diagram, (ii) a student's written interpretation of that diagram, and (iii) the teacher's normative RISN model of the diagram. The application constructs a RISN model of the student's interpretation, compares it with the teacher's model, and derives feedback for tailoring lessons. Initial trials show some promise that a more robust system could be built. However, issues such as consistency in constructing RISN models, transparency in the AI assistant's internal processes for building the system and improving the usability of the output remain to be addressed.

1 INTRODUCTION & VISION

Despite the ubiquity of diagrams, charts and graphs in education, formal training on their use is lacking. Thus, it is not surprising that, when students are formally evaluated in examinations, a variety of interpretations and misconceptions can surface (e.g., [6], [8]).

Our vision is to create a diagnostic tool for teachers that helps evaluate students' ability to interpret graphs and charts, independently of their understanding of the topic content. We aim to do this in a way that is grounded in cognitive theory of how external representations are interpreted. To achieve this, we use an AI assistant to help develop a pilot web application. This application is not a tutoring system (such as [7] or many LLM applications [9]), but a tool that supports educators to diagnose the quality of *students' graph interpretations*.

The cognitive theory we adopt is Cheng's cognitive *Representational Interpretative Structure Theory* (RIST) and its accompanying notation (RISN) [1, 3, 4, 5]. It proposes that an interpretation of an external representation is a hierarchical network comprised of four kinds of mental schemas: *Representation*, *R-scheme*, *R-dimension* and *R-symbol*. Figure 1 shows an example of a RISN model for interpreting a simple three bar Income by Age-profile chart. It contains: a *Representation* schema of the chart (Figure 1, E1), *R-schemes* that coordinate data, age and data points (E2, D3), *R-dimensions* for each of those dimensions (B4, E4, G4), and *R-symbols* for

specific data points (B6, D6, F6). Crucially, each schema coordinates information between a target concept (internal representation) and its graphical representation (external representation). Different interpretations create different schema networks. Stockdill et al. [10] identified recurring patterns of schemas in these network structures, called *idioms*.

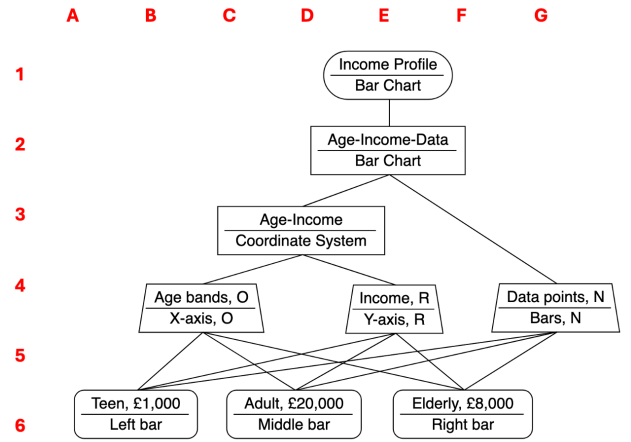


Figure 1. A simple example of a RISN model for a line graph. The red index grid was added for clarity.

RIST/N can be used to model interpretations of various diagrams [10], evaluate the usability of representations [3], study cognitive artefacts such as indexing systems [4], and observe changes in knowledge structures during learning [2]. However, it has not yet been applied in an educational context, as a key challenge is training people to create RISN models. Thus, in this work, we use a Large Language Model (LLM) as the underlying engine for automatic creation and comparison of RISN models. This is an aspect that has not been explored and would be beneficial to automate.

Our vision is as follows. We start by modelling a student's written interpretation as a hierarchical RIST/N network to identify omissions or misinterpretations of concepts and diagram relationships. An LLM is used: (a) to generate a RISN model of a student's interpretation; (b) to compare it to the teacher's target RISN model; and (c) to provide a report to the teacher of missing schemas and global and local misinterpretations. Figure 2 illustrates this pipeline. Here we will consider three questions about such a system:

- Q1. Can an application be created so that it can produce reasonable RISN models from a student's written description of a diagram?
- Q2. Can it successfully compare a student's RISN model with a teacher's target RISN model?

¹ Department of Informatics, University of Sussex, BN1 9QH, UK. Email: {g.garcia-garcia, p.c.h.cheng}@sussex.ac.uk.

Q3. Can it provide reliable, comprehensible and useful summaries of the differences between the student and the teacher models verbally and graphically?

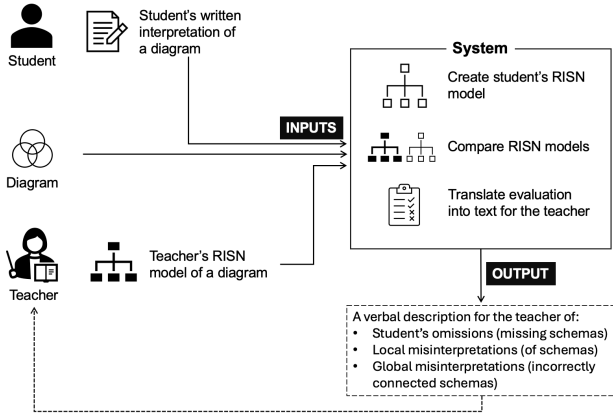


Figure 2. Pipeline for using RIST/N in an educational setting.

2 WEB APPLICATION

We developed a pilot web application with the assistance of Claude’s Sonnet 4.6. It was used both during the development of the web application and was integrated into the system for background processing. Published papers on RIST/N were fed to Claude, and it implemented the application in JavaScript and HTML.

The system was refined through multiple iterations. We provided examples of simple to complex line graphs (e.g., speed/distance graphs, Playfair’s trade balance line graph) and a variety of interpretative statements (including misconceptions such as the graph-as-picture misconception [6]). We reviewed the generated RISN models and reports for the teacher and prompted the AI assistant to modify the code when issues arose. To support development, we added an optional feature to generate synthetic interpretations of the diagrams at novice, intermediate and expert knowledge levels – this was used in addition to testing the system with actual users’ interpretations collected from previous research projects.

In use, the web application receives a diagram, a student’s written interpretation of a diagram, and the teacher’s normative RISN model (Figure 2, left). Internally, the application generates a student’s RISN model from the student’s interpretation, compares it with the teacher’s model, and produces a summary of the differences between them (Figure 2, top right), making calls to Claude as needed. The output is a report highlighting omitted concepts as well as global and local misinterpretations (Figure 2, bottom right).

3 SYSTEM REVIEW

We reviewed each of the key steps of the pipeline: examined the generated RISN models to ensure they fit with RISN’s rules, verified whether the student’s written interpretation matched the generated RISN model, and conducted heuristic evaluations of the generated teacher’s feedback. The following list describes some of the main issues.

Creating a RISN model from a student’s written interpretation. The creation of the student’s RISN model was inconsistent, as different models were generated from the same input. Despite this, most generated models were valid (i.e., well-constructed), and the validity of RISN models improved with minor refinements of the prompt regarding RISN construction rules. To address the production of inconsistent models, however, we may need to add rules regarding the generation of RISN models. For instance, we may need to specify that RISN schemas should be added in the order they appear in the written description, that schemas should be created statement by statement rather than allowing the LLM to infer relationships automatically, and to limit the topic knowledge that the LLM introduces in the creation of models.

Comparing a student’s and a teacher’s RISN models. Inspecting the Claude-written code, we found two methods were used to compare models: (1) matching strings of text from the schemas, and (2) sending the RISN models to Claude to perform structural comparisons. This caused inconsistencies in the output. For example, two strings of text could match, but if their schemas were in different locations within the student’s and the teacher’s models, then it was marked as a mismatch. Also, semantically equivalent strings were marked as different (e.g., ‘life stage’ and ‘age group’ would be marked as different). Although this could be remediated by providing further explicit instructions on how to compare models, this highlights the potential imprecision of using an AI assistant to process data in the background.

R-DIMENSION COMPARISON			
STATUS	GENERATED CONCEPT - GRAPHIC - SCALE	REFERENCE CONCEPT - GRAPHIC - SCALE	SCALE MATCH
matched +80%	average height vertical axis with numeric labels 0-200	Height D2	Yes: R-1 Ref: N-1
only generated	life stage horizontal axis with four category labels	—	O→N
only generated	country line colour (orange vs green)	—	N→N
only in reference	—	Age group D3	R→R
only in reference	—	Data record D4	N→N
only in reference	—	Record of the height of age group per region (D1, D2, D3, D4)	N→N

OMISSIONS	
•	The student omits the Region dimension (D1, scale N→N) as a distinct top-level schema separate from colour encoding.
•	The student does not include United Kingdom and Mexico as R-Symbols explicitly anchored to a Region dimension.
•	The student omits the Data record dimension (D4, scale N→N) entirely.
•	The student does not model the per-region sub-record split into D41 (UK) and D42 (MX).
•	The student omits the derived rate-of-growth R-Symbol computed via anchoring between Height-of-children and Height-of-teens.
•	The student does not use composite tuple coordinate notation (e.g. S11,S21,S31,S421) for individual data points.
•	The student omits the mereological decomposition of the overall data record into per-region sub-records.

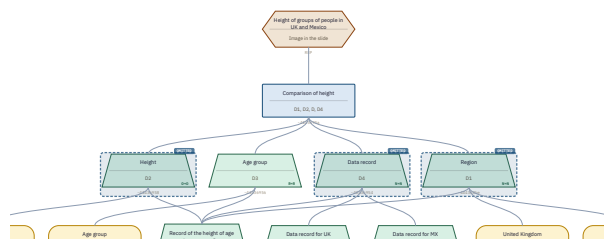


Figure 3. Snapshots of the feedback output for teachers. Formats from top to bottom: table, text summary, and annotated graph.

Output generated for the educator. Different formats were explored for presenting feedback for the teacher: (a) a table comparing the schemas and idioms from both the student’s and teacher’s models; (b) a text summary of omissions, local and global misinterpretations; and (c) the teacher’s RISN model

annotated with missing concepts, local and global misinterpretations. See Figure 3. In terms of usability, the table presentation provided the best way to quickly review missing concepts, while the text summary was too dense to be practical, and the annotated graph was inconsistent with the text output.

4 DISCUSSION & FUTURE WORK

In this exploratory study, we developed a pilot web application that uses an LLM to create and compare RISN models for supporting educators in understanding students' mental structures of their diagrammatic interpretations. In particular, it focuses on supporting educators in identifying whether students have omitted relevant concepts and their relationships, including both local and global misinterpretations.

Consider the three questions posed. For (Q1), we found that the system can generate reasonable RISN models, but identical inputs can produce different models. Addressing this requires clearer rules for deriving RISN models along with constraints on additional topic knowledge introduced by the LLM.

For (Q2), it was found that the application could compare the student's and teacher's RISN models. However, Claude applied two different comparison methods, which resulted in inconsistent feedback. Our intention was not to use different methods; however, this highlights that (a) prompt design requires careful attention, including the explicit specification of implicit logical constraints (e.g., using the same comparison method across all stages), and (b) code generated by agents must be reviewed.

On the other hand, this issue highlighted the need to develop methods for comparing and identifying (equivalent) network structures within RISN models. Further investigation would then need to use a wider variety of diagrams and students' (mis)interpretations to identify which method is most accurate. Once an effective method is developed, Claude could be instructed to include it as part of the system.

(Q3) concerns the reliability, comprehensibility and usefulness of the summaries of the differences between the student and the teacher models, verbally and graphically. This pilot study represents an initial step towards addressing this question. Of the various reporting formats explored, tables appeared to be the most useful, as they provided a quick visual way to highlight missing concepts. Nevertheless, alternative formats that offer additional value for the teacher may need to be explored. For example, feedback could be more concise and focused through brief three-point summaries of critical issues.

This study revealed that there are ample opportunities for future work. For example:

- We presented the example using a single graph interpretation as an input. However, the broader vision is for the system to collect graph interpretations from multiple students to enable the teacher to provide immediate feedback during the class. This is similar to using a short poll during a class or workshop, where participants' responses help guide subsequent activities.
- The evaluation of this system should include feedback from educators on the reliability and practicality of recommendations, as well as its potential for implementation in classroom settings.
- The system should be tested more extensively using a wide variety of diagrams, students' interpretations (from

novices to experts), and types (correct, incorrect, partial interpretations, and misconceptions). These should be collected from real people, as the extent to which synthetic data may have influenced the generation of RISN models remains unclear.

- Finally, an alternative scenario could be explored in which students *also* receive feedback on how their interpretations differ from the teacher's expectations. Although this approach will require consideration of other factors (e.g., students' metacognitive skills, confidence and engagement on the topic), it could provide valuable opportunities for students to generate explanations and deepen their understanding of the use of graphs and diagrams.

REFERENCES

- [1] Cheng, P. C-H. (2020). A sketch of a theory and modelling notation for elucidating the structure of representations. In A. Pietarinen, P. Chapman, L. Bosveld-de Smet, V. Giardino, J. Corter, & S. Linker (Eds.), *Lecture Notes in Computer Science: Vol. 12169. Diagrammatic Representation and Inference* (pp. 93-109). Cham: Springer.
- [2] Cheng, P. C-H., & Garcia Garcia, G. (2025). What and how schema networks are acquired during the learning of line graphs: Modelling using representational systems theory. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47.
- [3] Cheng, P. C-H., Garcia Garcia, G., Raggi, D., & Jamnik, M. (2024). A human information processing theory of the interpretation of visualizations: demonstrating its utility. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-14). Association for Computing Machinery, New York, NY, USA.
- [4] Cheng, P. C-H., Garcia Garcia, G., Raggi, D., & Jamnik, M. (2024). Index systems: Enumerating their forms and explaining their diversity with representational interpretive structure theory. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46.
- [5] Cheng, P. C-H., Stockdill, A., Garcia Garcia, G., Raggi, D., & Jamnik, M. (2022). Representational interpretive structure: theory and notation. In V. Giardino, S. Linker, R. Burns, F. Bellucci, J.-M. Boucheix, & P. Viana, (Eds.), *Lecture Notes in Computer Science: Vol. 13462. Diagrammatic Representation and Inference* (pp. 54-69). Springer, Cham.
- [6] Garcia Garcia, G. & Cox, R. (2009). Children who Interpret Graphs as Pictures. In V. Dimitrova, R. Mizoguchi, B. du Boulay & A.C. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education, AIED 2009* (pp. 695-697). IOS Press.
- [7] García-Méndez, S., de Arriba-Pérez, F. & Somoza-López, M.d.C. A. (2025). Review on the use of large language models as virtual tutors. *Science & Education*, 34, 877-892.
- [8] Kerslake, D. (1977). The understanding of graphs. *Mathematics in School*, 6(2), 22-25.
- [9] Lyu, W., Wang, Y., Chung, T., Sun, Y., & Zhang, Y. (2024). Evaluating the effectiveness of LLMs in introductory computer science education: A semester-long field study. *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24)* (pp. 63-74). Association for Computing Machinery, New York, NY, USA.
- [10] Stockdill, A., Garcia Garcia, G., Cheng, P. C-H., Raggi, D., & Jamnik, M. (2022). Cognitive modeling of interpretations of representations. In J. Macbeth, L. Gilpin, & M. T. Cox (Eds.), *Proceedings of the Tenth Annual Conference on Advances in Cognitive Systems* (pp. 36-56). Wright State University, Dayton, OH.

Flow-Centred Design: Rethinking Accessibility as an Experiential Quality in Human-Computer Interaction

Daniel Hajas¹

Abstract. Accessibility standards and assistive technologies continue to improve social inclusion by enabling disabled people to access places, information, and services. However, accessibility in Human-Computer Interaction remains primarily evaluated through task completion, compliance, and usability, often overlooking the experiential quality of interaction once access has been achieved. This position paper proposes *flow* as a complementary lens for reframing accessibility beyond access. Drawing on flow theory from psychology, I argue that accessible systems should not only enable participation but also support conditions for *sustained engagement*, *guided exploration*, and *skill development*. To operationalise this notion, I introduce *Flow-Centred Design (FCD)* as a conceptual framework linking accessibility evaluation with interaction design practice. FCD reframes accessibility as an experiential quality of interaction, grounded in design principles derived from conditions associated with flow experiences. The framework also introduces measures for evaluating the experiential quality and extent of accessibility across different interaction contexts. While the discussion focuses primarily on the digital inclusion of blind people, the proposed framework may be applicable to broader forms of disability across both digital and physical accessibility contexts. Through examples from accessible data visualisation research, I illustrate how interaction techniques supporting customisation, structured exploration, perceptual congruence, and multimodal collaboration begin to move accessibility beyond task completion toward richer, more engaging, and growth-oriented experiences.

1 INTRODUCTION

Over the past several decades, disability rights advocates have fought to secure equal participation in education, employment, and public life [4, 1, 2, 17]. These civil rights movements, alongside technological innovation, have led to major advances in the development and adoption of accessibility standards, assistive technologies, and inclusive design practices. In digital environments, accessibility guidelines and compliance-based evaluation have become central mechanisms for ensuring that disabled people can access information and services. These developments represent an important and necessary foundation for digital inclusion.

However, accessibility is often framed in terms of whether a task can be completed at all. In practice, this framing can obscure an important question: *what is the quality of the experience once access is achieved?* For many blind users, interacting with digital systems involves navigating layers of friction [12]. This often takes the form of long, linear navigation through poorly structured interfaces, as well

as the need to plan and mentally simulate interaction steps in advance because systems behave unpredictably. Over time, the constant problem-solving and adaptation required to cope with these design shortcomings generates emotional and cognitive fatigue that makes it difficult to remain immersed in the activity itself. In other words, when interaction requires constant compensation for inaccessible design, part of the user's attention and energy is spent overcoming barriers rather than pursuing meaningful goals. Consequently, blind people may perceive digital interaction as a more negative experience than their sighted peers, even when they are able to complete the same task. Yet, accessibility lacks a clear conceptual vocabulary for specifying and evaluating the experiential quality of interaction once baseline access has been achieved.

In response, this paper proposes *flow* as a complementary lens for evaluating accessibility in Human-Computer Interaction. Typically, flow is described in terms of the *optimal experience*. According to Mihály Csíkszentmihályi [7], flow occurs when “one’s skills are adequate to cope with the challenges at hand, in a goal-directed, rule-bound action system that provides clear clues as to how one is performing.” Drawing on flow theory from positive psychology, I argue that accessibility should not only enable participation, but also support conditions for meaningful engagement through *sustained engagement*, *guided exploration*, and *skill development*. From this perspective, accessible systems are those that create conditions for flow experiences for blind people and, respectively, for other disabled individuals.

To make a tangible contribution, the paper introduces *Flow-Centred Design (FCD)* as a conceptual framework for specifying and evaluating the experiential quality of accessibility. FCD proposes a set of design principles associated with conditions for flow and introduces families of measures concerned with *customisability and adaptability*, *exploratory structure and navigation*, and *learning and transferable skill development*. Accessible data visualisation is then used as a case study to explore how these conditions and measures may be instantiated in practice. Shifting the attention from minimal access toward the experiential quality of interaction, I argue that evaluating accessibility through the lens of flow can support more engaging, confidence-building, and growth-oriented forms of participation.

1.1 Positionality statement

I write this paper from the perspective of a blind individual, who identifies with both research and practitioner roles, working at the intersection of digital accessibility, Human-Computer Interaction, and disability innovation. My professional practice includes commentary on inclusive design in action, development of assistive technology prototypes, and accessibility testing as a consultant. As a scholar,

¹ University College London, England, email: d.hajas@ucl.ac.uk

I engage in co-design with subject-matter experts to explore new interaction approaches in accessibility, and I rely on collaborative tools for the dissemination of research findings. While my perspective is shaped by lived experience navigating digital systems through touch and text-to-speech technology alone, I am far from being a power user of any particular assistive technology. In investigating flow in accessible interaction, insights from expert users — analogous to elite performers in other domains — may reveal patterns of practice and expertise that extend beyond my own experience. As such, I intend this paper as a contribution to an ongoing interdisciplinary conversation about what access is, and what it is not.

2 FROM ACCESS TO FLOW: SPECIFYING AND MEASURING THE EXTENT OF ACCESSIBILITY AS AN EXPERIENTIAL QUALITY

In this section, I argue that accessibility should be reconsidered through the lens of *flow* — an optimal experiential state characterised by focused engagement, perceived control, and intrinsic reward[6]. Describing and building on characteristics of flow from existing literature in positive psychology, and their growing relevance within Human-Computer Interaction (HCI), I introduce *Flow-Centred Design (FCD)* as a framework for conceptualising accessibility as an experiential quality rather than solely a condition of access. Through FCD, this section pursues two aims: first, to specify the experiential quality of accessibility through a set of design principles derived from characteristics of flow experiences; and second, to propose directions for measuring the extent of accessibility through evaluative measures aligned with these principles.

2.1 Access is not enough

Research and practice in digital accessibility have traditionally focused on removing barriers to interaction between disabled people and digital content. Accessibility evaluation tools and frameworks typically ask whether systems comply with guidelines, and whether users can complete tasks, perceive content, or navigate interface elements using assistive technologies[22]. These assessment criteria are essential for establishing conditions for baseline inclusion. However, they primarily measure possibility rather than experiential quality and neglect aspects of experience design[10] and developmental flourishing[20]. As a result, a system may technically satisfy accessibility requirements while still imposing significant friction in interaction; requiring sustained cognitive effort, fragmented navigation strategies, and external tools for gaining proficiency.

Consider Silvia, a blind professional using a workplace chat application that is technically compatible with a screen reader and therefore considered accessible under conventional criteria. They can send messages, review conversations, and navigate channels independently without assistance. However, the interaction may still be effortful, fragmented, and cognitively demanding in ways that diminish sustained use. For example, when reviewing messages sent earlier in a conversation, screen reader focus may repeatedly jump back to the message input field whenever a new message arrives, interrupting concentration and forcing the user to reconstruct conversational context. In threaded replies, the interface may announce quoted messages before the new response itself, requiring additional effort to infer what information is novel. Similarly, metadata such as the sender’s name may be repeatedly announced in one-to-one conversations where identity is already known, increasing verbosity and

interaction fatigue. Although such an application may satisfy accessibility requirements in the sense that interaction is possible, the experience itself may remain frustrating, tiring, and difficult to sustain. Navigation becomes fragmented, contextual awareness is repeatedly disrupted, and users may seek to complete interactions as quickly as possible rather than remain meaningfully engaged. In this sense, accessibility alone does not guarantee an interaction environment conducive to focused, confident, or rewarding participation.

2.2 Accessibility beyond access

I argue that accessibility evaluation should extend beyond barrier removal, to consider whether interaction environments support conditions for the state of flow. When an interaction is accessible and the optimal experience is achieved — that is, the user is in flow — users can direct their attention toward pursuing clear goals. In doing so, they can displace self-consciousness with conscious effort, feel and report positive emotions, and become more proficient, without the distractions of managing the interface itself. However, flow experience is inherently subjective and context-dependent. It arises from the dynamic relationship between a person’s skills, goals, and the challenges presented by a task [6].

Although flow experiences are difficult to design for directly, prior work in positive psychology identifies a set of recurring characteristics associated with entering and sustaining flow states[6, 19]. These characteristics describe the experiential conditions under which individuals become deeply absorbed in meaningful activity, balancing challenge, attention, feedback, and perceived control. While originally formulated in domains such as play, sport, work, and creative practice, these conditions offer a useful conceptual foundation for reconsidering accessibility as an experiential quality of interaction. Several conditions are particularly relevant to accessibility contexts[19]:

1. Perceived challenges and opportunities for action should stretch skills and abilities, though not excessively.
2. Goals should be relatively short-term (proximal), specific, and detailed.
3. Immediate and actionable feedback should be available to promote continuing progress.
4. Concentration should be focused and intense.
5. Action and awareness should merge.
6. Reflective self-consciousness should fade or be lost.
7. A sense of control over actions and responses to changing situations should increase.
8. Time may become distorted, often appearing to pass more quickly than in reality.
9. Experiences should feel intrinsically rewarding, such that the activity itself justifies the effort involved.

In accessibility contexts, this variability of flow experiences is particularly pronounced. Blind users employ diverse interaction strategies and assistive technologies, often shaped by individual workflows and long-term adaptation to specific tools [21]. To operationalise flow as a measure of accessibility, I propose *Flow-Centred Design (FCD)* as a conceptual framework that links accessibility evaluation with experience design practice. Rather than prescribing a single optimal interaction technique, FCD treats flow as an evaluative heuristic — a way of identifying where friction disrupts engagement, where interaction demands exceed available skills, and where opportunities exist to support more perceptually congruent interaction [14]. In this

framework, accessibility is reframed not only as a condition of access, but as an experiential quality that may be supported to varying extents across different systems, users, and contexts.

Through FCD, two related aims are pursued. First, the experiential quality of accessibility is specified through a set of design principles derived from characteristics of flow experiences and adapted to accessibility contexts. Second, the extent of accessibility is considered through possible measures aligned with these principles, offering an initial conceptual vocabulary for evaluating how accessible interactions are experienced in practice.

2.3 Design principles for specifying the experiential quality of accessibility

The nine conditions outlined above describe characteristics associated with entering and sustaining flow experiences, but they are not immediately actionable as design goals for accessibility. To operationalise these characteristics in accessibility contexts, I synthesise and transpose them into three higher-level design principles that reflect recurring experiential qualities of interaction. Rather than reproducing the original psychological constructs directly, these principles adapt conditions of flow to concerns central to accessible interaction, such as reducing friction, maintaining orientation, supporting confidence, and enabling progressive growth in competence. Mapping the conditions for flow experiences to these principles, I propose that accessible systems should support *sustained engagement*, *guided exploration*, and *skill development* in meaningful activities, including learning, collaborative work, and creative expression.

Sustained engagement: Accessible systems should support the conditions necessary for maintaining focused and uninterrupted participation in meaningful activity. This principle aligns most strongly with conditions related to concentration, immersion, and intrinsic motivation. Specifically, flow requires *focused and intense concentration* (4), the *merging of action and awareness* (5), and the *fading of reflective self-consciousness* (6), all of which depend on interactions becoming sufficiently effortless that users can attend to the activity rather than the interface itself. In accessibility contexts, excessive friction — such as fragmented navigation, high cognitive load, or inaccessible controls — redirects attention toward managing the interaction, disrupting engagement. Sustained engagement also relates to *intrinsic reward* (9), whereby interaction feels meaningful and worthwhile beyond task completion alone. Flow-centred accessible systems should therefore minimise avoidable attentional disruption and support prolonged, rewarding engagement in activities such as reading, communication, creative work, or learning.

Guided exploration: Accessible systems should provide structure and feedback that support confident navigation through unfamiliar or evolving interaction spaces. This principle maps most closely to conditions concerning *proximal, specific goals* (2), *immediate and actionable feedback* (3), and an increased *sense of control* (7). Flow emerges when users can orient themselves toward clear objectives, understand the consequences of their actions, and adapt to changing circumstances without uncertainty or disorientation. In accessibility contexts, guided exploration may involve clear navigation pathways, progressive disclosure of complexity, predictable interface behaviour, and timely multimodal feedback that supports orientation and decision-making. Rather than prescribing a single interaction strategy, systems should enable users to explore while preserving confidence that progress is understand-

able, reversible, and responsive to individual workflows and assistive technologies.

Skill development: Accessible systems should support gradual increases in competence by balancing challenge with the user's evolving abilities. This principle aligns primarily with the requirement that *challenges and opportunities for action stretch skills without exceeding them* (1), creating conditions for growth rather than frustration or boredom. Over time, interactions should scaffold progression from effortful engagement toward fluency, allowing users to build proficiency with tools, workflows, and interaction strategies. Skill development also relates to the altered *experience of time* (8), often associated with deep engagement in progressively mastered activities, where sustained practice feels absorbing rather than effortful. In accessibility contexts, systems should support learning through adaptable levels of complexity, opportunities for mastery, and interactions that evolve alongside user expertise, enabling long-term flourishing rather than merely short-term access.

To illustrate these principles in practice, consider *Sophie*, a blind professional completing her weekly grocery shopping using a screen reader-accessible mobile application. Her primary cognitive effort should be directed toward meaningful decisions: determining how much she can afford to spend that week, estimating what can fit in her refrigerator, planning meals before ingredients expire, and reasoning about which items are needed for breakfast, lunch, and dinner. A flow-centred accessible system would support *sustained engagement* by minimising friction in interaction, allowing Sophie to compare products through a clear list structure, quickly add items to her basket without losing browsing context, and remain focused on meal planning rather than interface management. *Guided exploration* would be supported through predictable navigation, actionable feedback, and clear system state awareness; for example, immediate confirmation when products are added to the basket, notifications of out-of-stock substitutions, updates on spending totals, and reminders of the number of items selected. Over time, *skill development* would emerge as the system adapts to Sophie's preferences, enabling her to shop more efficiently, discover cost-saving alternatives, reuse prior shopping patterns, and gradually optimise her decisions around saved favourites and meals. In this example, accessibility is not only reflected in whether shopping is technically possible, but in the extent to which the interaction enables focused, confident, and rewarding participation in an everyday activity.

2.4 Measuring the extent of accessibility

If accessibility is understood as an experiential quality, then accessibility evaluation must also move beyond determining whether interaction is technically possible. The second contribution of Flow-Centred Design is therefore to suggest how the *extent of accessibility* might be evaluated in relation to the design principles introduced above: *sustained engagement*, *guided exploration*, and *skill development*. Rather than treating these principles as abstract aspirations, FCD proposes that they may serve as evaluative lenses through which the experiential quality of accessibility can be examined. This does not imply that flow can be measured simply or universally. Indeed, prior work identifies substantial variation in how flow is measured, with no single instrument adequately determining whether a person is in or out of flow [16]. For this reason, the measures proposed below should be understood as provisional, non-exhaustive families of evaluation for assessing how accessible interactions are experienced in practice.

Sustained engagement (through interaction friction and adaptation):ceptual vocabulary for examining the extent to which interactions support *sustained engagement*, *guided exploration*, and *skill development*. By reframing accessibility as an experiential quality, Flow-Centred Design offers a way of identifying where interaction supports or disrupts opportunities for meaningful participation and developmental flourishing. In the following section, accessible data visualisation is used as a case study to examine how these design principles and associated measures may be applied in practice, demonstrating both the strengths and limitations of contemporary interaction techniques.

The extent of accessibility may be evaluated by examining the degree to which interactions reduce unnecessary friction and adapt to users' abilities, preferences, goals, and evolving expertise. In Flow-Centred Design, accessible systems should support sustained engagement by maintaining an appropriate balance between challenge and skill, recognising that interaction strategies vary considerably across users and contexts. Measures may therefore include the *degree of customisability*, such as what content is spoken, how verbose this content is, or when it is read in sequence in the context of screen reader interaction[11]; the *extent of personalisation*, including adaptive support for different expertise levels, mediated by semantic scaffolding[25]; the *degree of user control*, reflected in whether interactions can be modified or tailored according to individual workflows; as well as the *extent of interactional friction*, reflected in interruptions to concentration, repeated corrective actions, or reliance on compensatory strategies. Rather than assuming a single optimal interaction pattern, greater accessibility may be reflected in systems that support multiple ways of engaging with information while reducing avoidable effort and adapting interaction demands over time.

Guided exploration (through orientation and navigational support):

The extent of accessibility may also be reflected in how effectively systems support orientation and confident exploration through interaction. Accessible interactions should provide meaningful structures that help users orient themselves, navigate confidently, and progressively deepen engagement with information. Measures may include the *mode of interaction structure*, such as whether information is organised through lists, tables, trees, graphs, or other representations suited to different tasks and reasoning strategies; the *diversity of navigational pathways*, including sequential, structural, spatial, or targeted navigation methods[23]; as well as the fluidity, reversibility, predictability, and contextual continuity of movement between states[14]. Systems that reduce disorientation, preserve context, and support independent exploration may therefore be perceived as more accessible than those requiring fragmented or effortful navigation.

Skill development (through progressive competence and transfer):

Finally, the extent of accessibility may be evaluated through a system's ability to support increasing competence and transferable proficiency over time. Flow-Centred Design proposes that accessible systems should not only enable participation in the present moment, but also facilitate progression toward greater competence, confidence, and autonomy. Measures may include the *rate of skill acquisition*, reflected in reductions in support needs — for instance, observed through less frequent prompting of a large language model-based Q&A assistive system, reduced navigation effort, or improved task completion time; the *transferability of interaction patterns*, where skills learned in one context can be applied across representations or tasks[14, 24]; and the *development of user confidence and independence*, reflected through qualitative feedback or longitudinal engagement. From this perspective, accessibility extends beyond immediate usability to encompass whether systems help users become increasingly capable participants in meaningful activity over time.

The measures proposed here are intentionally provisional and should be understood as illustrative families of evaluation rather than settled or comprehensive metrics. Their purpose is not to prescribe universal indicators of accessibility, but to provide an initial con-

3 ACCESSIBLE DATA VISUALISATION AS A CASE STUDY

Accessible data visualisation provides a useful case study for examining how interaction design can move beyond basic accessibility toward more experiential forms of engagement. In recent years, research in accessible data visualisation has increasingly shifted from ensuring access to graphical information toward enabling blind and partially sighted users to actively explore, manipulate, and reason about data through alternative representations. This emerging body of work provides a useful lens for examining how the design principles and measures proposed by Flow-Centred Design may be instantiated in practice. The following examples are not presented as definitive validations of FCD, but as demonstrative cases that illustrate how the experiential quality of accessibility may be supported and evaluated across different forms of interaction.

3.1 Structured exploration through accessible navigation

Rich Screen Reader Experiences for Accessible Data Visualisation [23] proposes a structured approach to screen reader interaction that moves beyond static textual descriptions of charts. Rather than presenting data visualisations as linear summaries, the system organises information hierarchically, enabling users to navigate between overview and detail, inspect subsets of data, and pursue analytical questions through structured exploration. This interaction model transforms screen reader use from passive information consumption toward active reasoning, allowing users to query and compare information according to their goals rather than relying on fixed descriptions. In doing so, the work reframes accessible visualisation as an exploratory process in which users can progressively deepen their understanding of a dataset.

From the perspective of Flow-Centred Design, this work aligns most strongly with the principle of *guided exploration*. Rather than overwhelming users with a single monolithic description, the hierarchical organisation supports gradual orientation and movement through the information space. Users are able to establish proximal goals, receive immediate feedback through interaction, and maintain a sense of control over their exploration, corresponding to several conditions associated with flow. The system also indirectly supports *sustained engagement*, as reducing navigational friction enables attention to remain directed toward reasoning about data rather than managing the interface itself. At the same time, reliance on hierarchical traversal may introduce limitations for some users. If users' mental models do not align with the imposed tree structure, or if interaction conventions differ from established screen reader workflows, exploration may become fragmented or cognitively effortful. For example, a user of Olli expected the tree structure to expand horizontally, similarly to a directory tree in Windows File Explorer, and

not vertically as it is currently implemented[3]. This suggests a possible tension between providing structured guidance and preserving low-friction, user-led exploration.

This work also helps contextualise how the extent of accessibility may be evaluated through the proposed family of evaluation related to *orientation and navigational support*. First, the system demonstrates one possible *mode of interaction structure*: a tree hierarchy that organises information into progressively deeper levels of detail. While this structure supports overview-to-detail exploration, later work demonstrates that alternative structures, such as graphs[8] or hypergraphs[14], may support different forms of reasoning. Second, the system expands the *diversity of navigational pathways* available to users, supporting structured traversal, spatial movement, and targeted search between levels of abstraction. From an FCD point of view, accessibility may therefore be understood not only in terms of whether a visualisation can be accessed, but in how richly users can navigate, orient themselves, and independently explore information.

3.2 Adaptable interaction through customisation

“Customization is Key”: *Reconfigurable Content Tokens for Accessible Data Visualisations* [11] extends accessible screen reader interaction by enabling users to actively reconfigure how information is presented. Rather than assuming a fixed description format, the system introduces configurable content tokens that can be permanently reordered, temporarily prioritised, and toggled or adjusted according to user preference and analytical goals. Thus, users are able to modify verbosity levels, foreground information relevant to a particular task, and tailor descriptions to suit both novice and expert workflows. In doing so, the work highlights a central challenge in accessibility: interactions that are effective for one user, context, or level of expertise may impose unnecessary friction for another.

This work aligns closely with the principles of *sustained engagement* and *guided exploration*, while also supporting *skill development* over time. By enabling users to adapt the interaction according to their preferences and expertise, the system helps maintain an appropriate balance between challenge and skill — a core condition of flow. For novice users, richer descriptions and greater guidance may reduce uncertainty and support orientation, whereas experienced users may benefit from concise, lower-friction interactions that enable faster reasoning. The ability to progressively modify interaction density also supports long-term proficiency, allowing workflows to evolve as users gain confidence and expertise. However, the act of customisation may itself introduce friction. Configuring verbosity, prioritising content tokens, or repeatedly adjusting interaction settings may interrupt concentration and temporarily disrupt sustained engagement. This highlights a potential tradeoff in Flow-Centred Design between adaptability and interactional continuity, particularly when configuration overhead outweighs experiential benefits.

Evaluating the accessibility, one could argue that the interface is fully accessible — without further discussion, since the fundamental interaction builds on the Olli interface[3]. However, this work strongly illustrates how the extent of accessibility may fall on an experiential spectrum, differentiated through the family of measures related to *interaction friction and adaptation*. Suitable measures may include the *degree of customisability*, reflected in the number and variety of ways content tokens can be reordered or configured; the *degree of user control*, reflected in the ability to modify verbosity, prioritise information, or adjust interaction pathways; and the *extent of personalisation*, reflected in how effectively the system accommodates differing expertise levels and goals. The work also reinforces

the relationship between adaptation and guided exploration, as adjustable structures and verbosity settings can influence how confidently and independently users navigate information spaces. Accessibility may therefore be understood not only as access to content, but as the extent to which interaction can be adapted to sustain meaningful and efficient engagement.

3.3 Transferable interaction through perceptual congruence

The concept of *perceptual congruence*, introduced through *Benthic: Perceptually Congruent Structures for Accessible Charts and Diagrams* [14], extends accessible visualisation by focusing not only on whether information can be accessed, but on how interaction structures correspond to the organisational logic of graphical representations. Rather than flattening charts and diagrams into linear descriptions or relying solely on hierarchical structures, Benthic introduces a hypergraph-based interaction model that preserves both hierarchical and relational groupings. This approach enables blind users to traverse representations fluidly, move between overlapping contexts, and reorganise exploration according to analytical goals. In doing so, the system supports forms of navigation and reasoning that more closely parallel the flexibility of sighted interaction, allowing users to develop richer mental models of diagrams and visualisations.

Here, perceptual congruence aligns strongly with the principles of *guided exploration* and *skill development*. The system supports guided exploration by enabling users to navigate complex information spaces in ways that are predictable, reversible, and responsive to changing goals. Rather than following fixed traversal paths, users are able to flexibly reorganise exploration, compare alternative perspectives, and recover from navigational detours without losing context. At the same time, Benthic supports transferable skill development through consistency in interaction logic across different graphical representations. By preserving underlying structural relationships, users can increasingly rely on learned reasoning strategies rather than repeatedly adapting to new interaction conventions, allowing skill acquisition to accumulate over time. Nevertheless, richer interaction structures may also increase cognitive demands, particularly for novice users unfamiliar with graph-based navigation or overlapping relational contexts. While perceptual congruence may support transferable reasoning strategies over time, early interactions may require greater effort to establish orientation and develop effective mental models.

Benthic illustrates how the extent of accessibility may be contextualised through both *orientation and navigational support* and *progressive competence and transfer* measures. First, the work expands the possible *mode of interaction structure* beyond tree hierarchies toward graph and hypergraph representations, demonstrating that accessibility may depend not only on whether structure is present, but on how appropriately that structure supports reasoning about relationships. Second, the system broadens the *diversity of navigational pathways*, enabling lateral, hierarchical, and reversible traversal, as well as flexible regrouping of information around user goals. Finally, perceptual congruence introduces a basis for evaluating the *transferability of interaction patterns*, reflected in whether users can apply familiar navigation strategies and mental models across domains such as charts, diagrams, or multimodal representations.

3.4 Multimodal and adaptive interaction ecologies

Umwelt: Accessible Structured Editing of Multimodal Data Representations [24] proposes a multimodal environment for creating and interacting with data representations in which visual, textual, and auditory modalities are treated as co-equal rather than hierarchical outputs. Instead of positioning accessible representations as secondary translations of visual artefacts, Umwelt derives multiple representations from a shared underlying data model, allowing users to fluidly move between modalities according to task demands and personal preferences. Users are able to independently author, modify, and reconfigure multimodal analytical workflows, switching between overview and detail while preserving context across representations. Consequently, Umwelt reframes accessibility as an ecosystem of interaction possibilities rather than a single optimised interface.

Umwelt represents perhaps the clearest illustration of how the three design principles may converge in practice. The system supports *guided exploration* through synchronised multimodal representations that preserve orientation and enable users to move flexibly between different ways of reasoning about the same information. It supports *sustained engagement* by allowing users to focus on analytical goals rather than on overcoming modality-specific barriers, reducing friction through complementary interaction pathways. Finally, Umwelt promotes learning and transferable *skill development* by enabling users to progressively refine workflows, build confidence across modalities, and adapt interaction strategies according to changing expertise or goals. Rather than prescribing a single ideal interaction, the system enables a dynamic balance between challenge and skill through user-led adaptation. At the same time, the availability of multiple modalities and interaction pathways may increase complexity and decision-making demands. Without sufficient guidance or Prior experience, users may exhibit uncertainty regarding which modality or workflow best supports a given analytical goal — a research topic often discussed, but not formally established.

Umwelt demonstrates how all three proposed families of evaluation may be contextualised within a single interaction environment. In terms of *interaction friction and adaptation*, the system supports extensive user control through runtime configuration, modality selection, and the ability to tailor representations according to analytical intent. Second, in case of *orientation and navigational support*, accessibility may be reflected in the diversity of available pathways, including movement between textual hierarchies, sonification, and visual representations while maintaining shared context. Finally, regarding *progressive competence and transfer*, Umwelt provides a basis for evaluating how users develop increasingly effective workflows, transfer reasoning strategies between modalities, and achieve greater independence in analytical tasks over time.

These case study examples suggest that accessible interaction design is increasingly moving beyond enabling task completion toward supporting richer forms of engagement. Across these systems, accessibility is not treated merely as the removal of barriers, but as the creation of conditions that support exploration, adaptation, and learning. Importantly, the case studies demonstrate both opportunities and tensions related to evaluating the extent of accessibility, which may be understood as a matter of an experiential spectrum, shaped by the structures available for exploration, the flexibility of interaction, and opportunities for developing transferable skills over time. While accessible data visualisation represents only one application domain, the broader implications of Flow-Centred Design may extend across accessibility research more widely, including education, assistive technologies, and access to physical or hybrid environ-

ments.

4 DISCUSSION

Flow has been widely applied in game design, educational technologies, creativity support systems, virtual and augmented reality, and user experience research as a framework for understanding engagement, immersion, motivation, and skill development. For example, game design has extensively adopted flow as a framework for balancing challenge and skill, structuring progression, and maintaining player engagement alongside broader usability and interaction principles[5]. Related ideas have also influenced areas such as educational technology, persuasive systems, and user experience design, where designers seek to sustain motivation, immersion, and intrinsic reward over time. In this sense, FCD does not introduce flow into HCI for the first time, but repositions it as a lens for accessibility evaluation, arguing that the experiential quality of accessible interaction deserves equal consideration alongside technical access and usability.

Importantly, Flow-Centred Design is not intended to replace established approaches such as User-Centred Design (UCD) or Human-Centred Design (HCD). Rather, FCD extends existing design frameworks by introducing flow as an additional evaluative lens for accessibility within the interaction design cycle. UCD and HCD have made substantial contributions to improving usability, reducing friction, and aligning technologies with human needs and contexts of use. However, accessible interaction is not solely determined by whether users can complete tasks or navigate interfaces, but also by the experiential quality of participation. Just as HCD expanded the scope of evaluation beyond usability toward broader user experience[15], FCD expands accessibility evaluation beyond access and usability — toward experience design and developmental flourishing[20]. FCD therefore complements existing approaches by asking not only whether interactions are usable and inclusive, but whether they support sustained engagement, reversible and low-stakes exploration, as well as long-term skill development in meaningful activities.

Emerging technologies create new opportunities for operationalising this approach. HCI has long recognised that no single interface can accommodate the diverse needs, preferences, and abilities of all users[18], motivating earlier movements in customisation, personalisation, end-user development, and malleable software[9, 13]. While such approaches have historically been constrained by implementation complexity and commercial incentives favouring standardised experiences, advances in artificial intelligence, multimodal interaction, and adaptive systems make personalised accessibility increasingly feasible. AI-assisted interfaces can dynamically adjust levels of guidance, feedback, modality, and task complexity in response to users' evolving skills, goals, and contexts of use. Although these capabilities must be implemented carefully to avoid over-automation, unpredictability, or diminished user agency, they offer promising mechanisms for maintaining an appropriate balance between challenge and skill over time — a central condition for enabling flow in accessible interaction.

4.1 Limitations and future work

Rather than proposing a definitive measurement framework, this paper offers an initial set of candidate measures grounded in the design principles of Flow-Centred Design. These measures are intended to

illustrate how the experiential quality of accessibility may be operationalised and evaluated in practice, while inviting further empirical refinement by the HCI and accessibility research communities. Future work may establish more consistent, valid, and robust approaches for evaluating the extent to which interactive systems support conditions for flow in accessible experiences.

A natural next step is methodological: empirically investigating how the design principles of Flow-Centred Design may be systematically observed and measured across different interaction contexts. For example, future work may identify behavioural indicators, qualitative accounts, and longitudinal patterns associated with *sustained engagement*, *guided exploration*, and *skill development*. In the context of *sustained engagement*, indicators may include reduced interruptions to concentration, fewer corrective actions, lower reliance on compensatory strategies, or qualitative reports of immersion, reduced fatigue, and intrinsic reward. For *guided exploration*, future studies may examine orientation, confidence, and contextual awareness during navigation, including the reversibility and predictability of interaction pathways, alongside qualitative accounts of perceived control and disorientation. Finally, *skill development* may be evaluated through changes in proficiency over time, such as reduced support needs, transferable interaction strategies across tasks, and increased confidence or autonomy in meaningful activities.

Such efforts would enable Flow-Centred Design to progress from a conceptual framework toward a more systematic methodology for evaluating accessibility as an experiential quality. Importantly, this work should remain sensitive to the contextual and subjective nature of flow experiences, recognising that meaningful accessibility may manifest differently across users, assistive technologies, and domains of activity.

5 CONCLUSION

Accessibility solutions have made digital inclusion and social participation for blind and other disabled people possible, but possibility alone is not enough. A system may technically satisfy accessibility requirements while still imposing significant friction, fragmented navigation, or barriers to developing proficiency over time. This paper argues that accessibility evaluation should extend beyond whether interaction is possible, to consider the experiential quality of interaction and the conditions that support meaningful engagement.

To address this gap, I proposed Flow-Centred Design as a conceptual framework for specifying and evaluating accessibility through the lens of flow. Rather than treating accessibility as binary compliance, FCD reframes it as a matter of an experiential spectrum, concerned with the extent to which interactions support sustained engagement, guided exploration, and skill development. In doing so, the paper introduced families of measures — including customisability and adaptability, exploratory structure and navigation, and learning and transferable skill development — as provisional ways of evaluating the experiential quality of accessibility.

Case studies in accessible data visualisation further suggest that these ideas are not purely speculative. Emerging interaction techniques increasingly support richer forms of exploration, personalisation, transferable interaction structures, and multimodal engagement that resonate closely with the aims of Flow-Centred Design. While the proposed measures remain intentionally provisional and require further empirical validation, they offer a starting point for future work on evaluating the extent of accessibility beyond access alone.

ACKNOWLEDGEMENTS

I would like to thank Jonathan Zong, who inspired me to express my scholarly opinion on Flow-Centred Design in a position paper, as a result of reading his excellent work on citational justice. I'm also grateful for everyone who shared their feedback, including Marianna Obrist, Emily Yates, and Jonathan Zong. Last but not least, my gratitude goes to Ronald Grau, who encouraged me to submit my own contribution to the symposium on Artificial Intelligence and Cognitive Science in Education, at AISB 2026.

REFERENCES

- [1] Colin Barnes and Geoff Mercer, *Disability*, Polity Press, Cambridge, 2003.
- [2] Len Barton, *Disability, Politics and the Struggle for Change*, David Fulton Publishers, London, 2001.
- [3] Matthew Blanco, Jonathan Zong, and Arvind Satyanarayan, 'Olli: An Extensible Visualization Library for Screen Reader Accessibility', in *IEEE VIS Posters*, (2022).
- [4] James I. Charlton, *Nothing About Us Without Us: Disability Oppression and Empowerment*, University of California Press, Berkeley, 1998.
- [5] Jenova Chen, 'Flow in games (and everything else)', *Commun. ACM*, **50**(4), 31–34, (April 2007).
- [6] Mihaly Csikszentmihalyi, *Finding Flow: The Psychology of Engagement with Everyday Life*, Basic Books, 1998.
- [7] Mihaly Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*, Harper & Row, 2009.
- [8] Frank Elavsky, Lucas Nadolskis, and Dominik Moritz, 'Data Navigator: An Accessibility-Centered Data Navigation Toolkit', *IEEE Transactions on Visualization and Computer Graphics*, 1–11, (2023).
- [9] Gerhard Fischer, 'Meta-design: Beyond user-centered and participatory design', in *HCI International*, (2003).
- [10] Giles Hamilton-Fletcher, Marianna Obrist, Phil Watten, Michele Mengucci, and Jamie Ward, "'i always wanted to see the night sky': Blind user preferences for sensory substitution devices", in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, p. 2162–2174, New York, NY, USA, (2016). Association for Computing Machinery.
- [11] Shuli Jones, Isabella Pedraza Piñeros, Daniel Hajas, Jonathan Zong, and Arvind Satyanarayan, "'Customization is Key": Reconfigurable Textual Tokens for Accessible Data Visualizations', in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, p. 14, Honolulu, HI, USA, (2024). ACM.
- [12] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey, 'What frustrates screen reader users on the web: A study of 100 blind users', *Int. J. Hum. Comput. Interaction*, **22**, 247–269, (05 2007).
- [13] Henry Lieberman, Fabio Paternò, Marcus Klann, and Volker Wulf, 'End-user development: An emerging paradigm', in *End User Development*, 1–8, Springer, (2006).
- [14] Catherine Mei, Josh Pollock, Daniel Hajas, Jonathan Zong, and Arvind Satyanarayan, 'Benthic: Perceptually congruent structures for accessible charts and diagrams', in *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '25, New York, NY, USA, (2025). Association for Computing Machinery.
- [15] Donald A. Norman, *The Design of Everyday Things: Revised and Expanded Edition*, Basic Books, New York, 2013.
- [16] C. Norsworthy, B. Jackson, J. A. Dimmock, and B. T. Johnson, 'Advancing our understanding of psychological flow: A scoping review of conceptualizations, measurements, and applications', *Psychological Bulletin*, **147**(8), 806–827, (2021).
- [17] Richard K. Scotch, *From Good Will to Civil Rights: Transforming Federal Disability Policy*, Temple University Press, Philadelphia, 1984.
- [18] Ben Shneiderman and Pattie Maes, 'Direct manipulation vs. interface agents', *Interactions*, **4**(6), 42–61, (November 1997).
- [19] C. R. Snyder and S. J. Lopez, *The Oxford handbook of positive psychology*, Oxford University Press, 2011.
- [20] Yuanyang (YY) Teng and Darren Gergle, 'Access is not enough: Toward developmental flourishing', in *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing*

- Systems*, CHI EA '26, New York, NY, USA, (2026). Association for Computing Machinery.
- [21] Lily M. Turkstra, Tanya Bhatia, Alexa Van Os, and Michael Beyeler. Assistive technology use in domestic activities by people who are blind, 2025.
 - [22] W3C. Web accessibility evaluation tools list.
 - [23] Jonathan Zong, Crystal Lee, Alan Lundgard, Jaemin Jang, Daniel Hajas, and Arvind Satyanarayan, 'Rich screen reader experiences for accessible data visualization', (2022).
 - [24] Jonathan Zong, Isabella Pedraza Piñeros, Ming Chen, Daniel Hajas, and Arvind Satyanarayan, 'Umwelt: Accessible structured editing of multimodal data representations', (2024).
 - [25] Jonathan Zong, Isabella Pedraza Pineros, Mengzhu Katie Chen, Daniel Hajas, and Arvind Satyanarayan. Semantic scaffolding: Augmenting textual structures with domain-specific groupings for accessible data exploration, 2025.

First steps towards novel adaptive systems for individualised music learning and assessment: An investigation of memory chunking in a piano chord production task

Christopher Luu¹ and Ronald R. Grau¹

Abstract. This paper presents a simple pilot study that uses cognitive behavioural measures to assess piano playing proficiency. The wider objective is to explore how digital game-based learning environments could be designed to adapt to learner behaviour in the musical domain. We believe it is possible to utilise cognitive behavioural metrics to not just assess aspects of musical proficiency but more importantly, to drive the selection and configuration of suitable learning sessions as well as teaching methods and content. With our approach, we aim to create a learning paradigm that adapts dynamically to a learner and avoids overly rigid instruction approaches that focus on the precise replication of musical patterns, which can make it difficult for learners of music to experience progress and stay motivated. Instead, we propose that cognitive behavioural metrics could be used to inform learning trajectories that give insight into learning progress through changes observed in the cognitive processes that underlie learning. In this paper, we present the results of a pilot experiment that compares the memory chunking signals obtained from participants of different musical proficiency who engaged in a musical chord reproduction task. The purpose of this experiment was to establish if results from previous research that targeted other domains of learning (e.g., maths, writing, drawing, etc.) are, in principle, transferable to the music domain. The results found indicate that this is likely the case.

1 INTRODUCTION

Pedagogies that assess music learners solely based on task accuracy promote a rote approach to learning - being only concerned with the precision of outputs rather than the indication of a deeper understanding of the target domain. For example, accuracy may be a suitable metric for classical pianists who seek to strictly maximise execution fidelity of a particular musical piece. In contrast, many jazz pianists will routinely use improvisation to enrich their performance, and these musicians must associate knowledge between various other aural and conceptual aspects to be proficient in their area. Accuracy cannot capture all elements of improvisation, nor of piano ability in general, as the ability to interpret music is also an indicator of proficiency and hence, there rarely is a single standard for comparison. A more cognitively informed digital learning environment could relate aspects of task performance to relevant cognitive functions, capable of constructing a higher-dimensional model of learner ability to

identify areas of weakness that may be limiting performance in other areas.

2 RELATED WORK

Interpreting behavioural measures to study mental processes is a long-established method of investigation in Cognitive Science, and a large body of work exists that has investigated memory chunking in the context of different tasks. Specifically, the pauses that are made by individuals when they are performing a task were found to coincide with information retrieval and processing activities, and the temporal patterns of pauses between actions can give insight into how information is organised in human memory. We also know that the size of memory chunks is related to our ability to retain different amounts of information, and that chunking structure also reflects the processes underlying working memory management and can indicate the strength of association between memory contents. [2][4].

Experts and novices structure information in working memory differently. Given the same task, experts are generally able to store more information per chunk than novices and related tasks induce less cognitive load [10][16]. The amount of cognitive load experienced can be estimated by measuring behavioural patterns, such as the timings of a student's actions. Student proficiency for a given task could therefore be measured by observing how their behavioural patterns compare with those typically observed in experts and novices [9][14]. Although there can be individual differences between people with equivalent degrees of proficiency, previous research has shown that similarities within their behavioural patterns can be extracted and analysed. For example, Cheng and Rojas-Anaya [6] were able to successfully categorise the mathematical competency of participants by analysing data about how they wrote down maths formulae: their writing actions, pen positions and timings, and the duration of pauses between drawn elements. Since playing piano involves a similar externalisation of intended actions through hand and finger movements, there is a distinct possibility here that chunking could contribute towards establishing a more comprehensive measurement of musical proficiency. We contend that one particularly promising way of obtaining such measures may be game-based learning environments, as these permit the collection of data not just over time, but also across different learning activities and contexts. Interpreting an individual learner's chunking pattern could help indicate when internal processing is happening, and point to areas of weakness in their musical skill that other methods might miss.

¹ Department of Informatics, University of Sussex, BN1 9QH, UK.
Email: {c.luu, r.r.grau}@sussex.ac.uk

Previous work has found strong correlations between proficiency and temporal behavioural metrics that were used to examine individual chunking structures in domains such as mathematics [5], natural language writing and comprehension [11], or computer programming [1]. Recent PhD theses have also examined chunking in the context of people reading musical notation [7] [12]. The pilot study presented in this paper aims to establish if people with different proficiency levels in piano playing might, in principle, exhibit chunking patterns that are consistent with the results previously found in those other domains.

3 APPROACH

The experiment discussed in this paper was conducted on Sussex AI Day in January 2026 at the University of Sussex. The event showcased current ongoing AI projects by research groups across the University. Using the Unity Game Engine [17], custom software was written in advance in C# to carry out the experimental trials and record relevant data during the event. There was a one hour window of time when attendees were able to volunteer to participate. This allowed for opportunistic sampling of participants who belonged to a range of different piano proficiency bands. Before starting the first trial, participants completed a pre-trial questionnaire related to their familiarity, confidence, and experience with the piano, which was used to categorise them each based on their general piano proficiency. The questionnaire had 12 questions, but to keep the categorisation process simple, only one question was used to determine an individual's proficiency group: "To what extent do you play piano?". This is because it would have been difficult to set weights for how each variable contributes to overall proficiency without prior testing. We gathered a total of 14 participants. The distribution of participants across categories is as follows: Complete Beginner (4), Hobbyist (6), Intermediate (3), Professional (1).

The experimental procedure involved 2 trials, preceded by a familiarisation trial, in which participants produced chords from memory immediately after viewing a stimulus of 15 chords, as shown in figure 1. Participants interfaced with the experimental trial using a standard 49-Key MIDI piano keyboard. The stimulus was hidden during chord production but could be viewed again upon request as often as necessary by holding down the space key on the laptop keyboard. Participants were instructed to use the same hand to both view the stimulus and to play chords with. This was to minimise task execution rehearsal while the stimulus was being viewed, since participants could not simultaneously get their hand into the correct position to play the next chord whilst viewing the stimulus.

The specific variables measured were determined by which behavioural metrics we hypothesised would correlate with piano proficiency. Average metric values were calculated for each participant, and then averaged again across proficiencies. They are as follows:

- M1)** Number of chords played between stimulus views
- M2)** Number of stimulus views per trial
- M3)** Stimulus view durations (seconds)
- M4)** Chord-playing durations (play periods) between stimulus views (seconds)
- M5)** Pause durations between chords (seconds)

This selection of metrics allows for the possibility of analysis to occur at a multiple-chunk scale (e.g. number of stimulus views), as well as at the level of the elements within individual chunks themselves (e.g. pauses between chords).

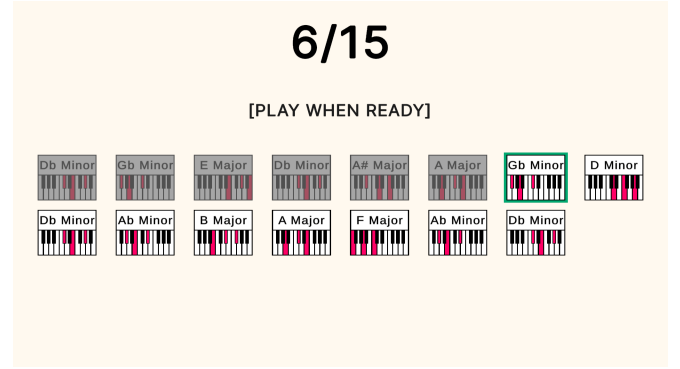


Figure 1. A screenshot from the experimental software, showing the chord progression stimulus for memorization (Trial 2). The next chord to be played is highlighted in green and the chords that have already been played are greyed out. "6/15" at the top refers to the number of chords the participant has played so far, out of 15. When the space-bar is released, the stimulus disappears from the screen.

4 RESULTS

As shown in figure 2, the results showed that higher proficiency was correlated with fewer stimulus views per trial, more chords played between stimulus views, and shorter pauses between chords. Correlations between stimulus view durations or play time durations with proficiency was unclear, which is discussed further in section 5.1. These results are consistent with previous work in other domains which find significant correlations between similar behavioural metrics and proficiency [1]. All descriptions of results are speculative, given that there was insufficient data to form any significant conclusions.

The data of 3 participants was excluded from these results for the following reasons:

1. Participant did not attempt to recreate the stimulus chords in any way and just pressed notes in ascending order.
2. Participant was too engaged in the exact scientific discussion before attempting the trial, which primed them to attempt to memorise far more chords (8-10) at a time than we believe they otherwise normally would have - skewing the results.
3. Participant did not seem to understand that they could re-view the stimulus upon request and attempted to memorise all chords in one try.

5 DISCUSSION & FUTURE WORK

5.1 Interpretation of Results

While further work will need to be done to properly control for these assumptions, this section discusses how the results for each metric (as shown in figure 2) may be explained by chunking processes.

- **M1 & M2):**

The average number of chords played between stimulus views measures how many chordal elements participants were able to recall. Since experts are able to incorporate more elements per chunk [10], we expect to see this metric increase as proficiency increases. For the same reason, we would expect to see the average number of times a participant chooses to view the stimulus

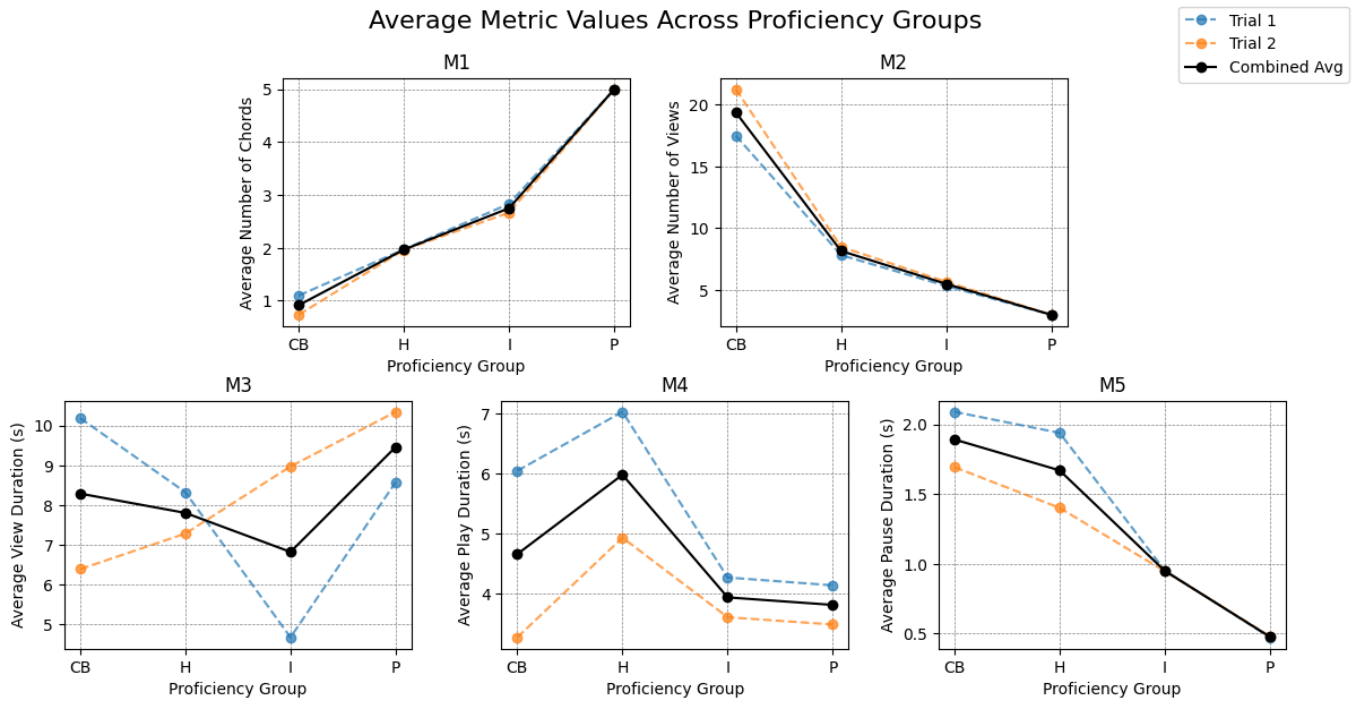


Figure 2. The observed results of all 5 metrics (M1-M5) gathered from behavioural data during both chord production memory trials. Proficiency groups are abbreviated as follows: Complete Beginner = CB; Hobbyist = H; Intermediate = I; Professional = P. Blue dashed lines represent trial 1 results; Orange dashed lines represent trial 2 results; Black solid lines represent combined trial results.

during trial decrease as proficiency increases, since a less efficient working memory will have to be refreshed more often. Both of these trends were observed in the results.

An experiment with similar methods tracked eye movements during a stimulus matching task to determine the conditions under which individuals favour online access - using the environment as its own model to complete the task as opposed to storing multiple elements of the stimulus in memory [3]. They found that participants would refer to the stimulus more if the cost of online access was less than the cost of memorisation. Looking at our study from this perspective, experts may produce more chords from memory because the size of their chunks result in the cost of memorisation being lower than the cost of requesting to view the stimulus. Understanding when individuals use the world as its own model and when they choose to rely on internal models from a chunking perspective may benefit our understanding of domain proficiency in a way that incorporates elements of 4E cognition.

- **M3):**

The average time taken to view the stimulus represents the amount of time it takes for a participant to fill their working memory with relevant chordal information to complete the memory task. The lack of a strong identifiable trend between view durations across proficiencies in the results may be explained by the fact that the time to perceive chunks is approximately constant - regardless of the number of elements within each chunk [4][1]. While analysing the data, we discovered that dividing the stimulus view durations by the number of chords that were memorised from that view provided a way of normalising this metric. Figure 3a shows that experts viewed the stimulus for less time per chord produced, indicating the presence of more efficient perceptual chunking of

chordal elements over novices.

- **M4):**

The average amount of time it takes for participants to produce the chords retained in working memory is likely independently affected by various factors. The participant's domain-specific motor ability will determine how fast they are able to play each chord, and therefore the amount of time to execute each full play period before having to review the stimulus. However, play duration will also be a function of how many chords were memorised as, *ceteris paribus*, it takes more time to play more chords. These confounding factors could explain the lack of a linear trend between play durations and proficiencies. Similarly to **M3**, normalising play period durations according to the number of chords played during that period reveals a decrease in the amount of time taken to play each chord as proficiency increases, as shown in figure 3b. Given that the processing of elements across chunk boundaries is shown to take more time than for elements within the same chunk [2][15], the faster production of chords from experts suggests the presence of larger chunks.

- **M5):**

The pause durations between each chord decrease as proficiency increases - which is a finding that is consistent with chunking theory. Experts likely hold whole chords as single homogeneous chunks that can be executed without further decomposition. Novices have to decompose each whole chord into sub-chunks to execute, which increases processing time, and therefore, pause duration.

While there is not enough data from this pilot study to provide any statistical significance, these results indicate that there may be

valid, operationalisable methods to identify piano proficiency based on analysing chunking structure. Furthermore, the promising trans-ferral to the musical domain of piano playing strengthens the theory behind previously identified works which utilise behavioural metrics to assess proficiency in other domains [5], [11], [1].

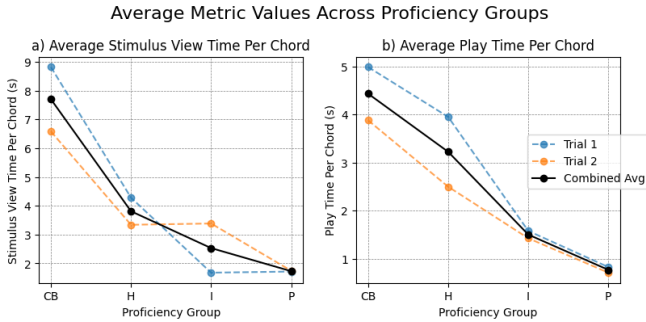


Figure 3. Metrics M3 and M4 normalised per chord: a) Average stimulus view time per chord - created by dividing M3 by M1. b) Average play time per chord - created by dividing M4 by M1. Proficiency groups are abbreviated as follows: Complete Beginner = CB; Hobbyist = H; Intermediate = I; Professional = P. Blue dashed lines represent trial 1 results; Orange dashed lines represent trial 2 results; Black solid lines represent combined trial results.

5.2 Limitations

5.2.1 Participant Data

Due to the small window of time that was assigned for the opportunistic sampling of participants, there were several issues regarding how data had to be collected, which prevented meaningful statistical analysis. The sample size was not large enough, nor was the distribution of participants across proficiency levels able to be controlled for. Factors that may have confounded with the observed results could also not be controlled for. This includes components that affect working memory access, such as age and various neurodivergent profiles, which are likely to influence individual task performance. Pre-trial proficiency evaluations could not be objectively determined, and relied on subjective answers from the given questionnaire. All observed trends are therefore merely indications of promising areas for future research. It is also worth noting that the public nature of the event may have introduced attentional challenges which distracted participants during trials. Future studies extending this methodology should attempt to control for these factors within a greater sample size. Work which evaluates whether the presented method can discriminate within a narrower range of proficiencies (e.g. between only intermediate players) would be particularly valuable for digital learning environment contexts.

5.2.2 Assessment Task Design

Assumptions were made about how the proposed metrics relate to internal cognitive mechanisms and chunking structures. For example, metrics **M1 & M2** assume that each new request to view the stimulus is representative of the fact that all of the chordal elements in working memory have been produced and that new elements need to be loaded in. However, there may be other factors affecting the rate of stimulus

viewing and chord production, such as self-confidence, which does not always align with proficiency (e.g. see the Dunning-Kruger effect [8]). Work needs to be done to thoroughly investigate the cognitive implications of each of the proposed metrics to increase their validity. This also includes more strict control over the difference in potential strategies employed by participants to ensure that behavioural metrics are cohesively measuring the same chunking phenomena.

The presented assessment method is ultimately only a proposed measure of the chord production element of piano playing, which is a subset of the skills required for an individual to be considered a skilled piano player. Further work will need to investigate how this kind of task can be transferred to other aspects of piano playing, such as sight-reading or scale production. As some of the proposed metrics rely on measuring temporal signals of behaviour, challenges may arise when applied to rhythmic aspects of piano playing which determine the exact timing of inputs required. All individuals who are able to match the target rhythm exactly will produce the same temporal signal, producing a ceiling effect on the data which would obfuscate the distinction between their levels of proficiency above a certain point. One way to circumvent this issue may be to gradually increase the difficulty of the stimulus, or the required tempo of the response - producing more variation in results between individuals by raising the ceiling effect. However, it is important to keep note that the more obscure the assessment method, the less likely the target of that assessment may align with the target skill - this is also a necessary direction for future research and is discussed further at the end of section 5.3.

5.2.3 Stimulus Design

Currently, no consideration has been made to investigate the effect that specific components of the stimulus have on resulting chunking signals. For example, experts may perform better when the stimulus content is presented according to common harmonic patterns, whereas this may not affect novice performance. There will also be variation as to the specific chords/ chord types that are easier to produce between individuals. Observing how idiosyncratic task performance changes in response to various stimulus components will indicate the strength of association that individual learners have with task-relevant musical elements. Research can then be directed towards effective ways to tailor the sequencing of learning content given these kinds of qualitative models of player proficiency. Human-Computer Interaction methods may also be able to reveal design implications for producing more reliable chunking signals by reducing the extraneous cognitive load [13] caused by visual elements of the stimulus and its layout.

5.3 Motor Skill vs. Conceptual Knowledge

The pilot study does not specifically distinguish between conceptual domain knowledge and motor ability. For example, metrics **M4 & M5** suggested that the average time taken to play each chord may be predictive of proficiency. However, these metrics are likely determined by both the conceptual knowledge of what notes are in each chord and the motor ability to efficiently configure the necessary body parts to physically play the chord - both could be viewed as distinct proficiencies. Given this, it is initially unclear how much of the chunking signal is distorted by complex interactions between motor execution and conceptual knowledge. Future work aimed at disentangling this interaction may be beneficial to clarify the validity of the proposed metrics, and for informing decisions related to

how digital learning environments use these metrics to adaptively sequence individualised learning content.

It is also worth considering the degree to which this distinction is necessary; studies from embodied cognition highlight the key involvement that sensorimotor loops have in working, implicit, and long-term memory (see [18] for a review). Since motor processes are seemingly involved throughout many key stages of cognitive processing, a more holistic approach to skill acquisition could involve purposefully blurring the lines between conceptual knowledge and motor skill. This is specifically pertinent to the domain of piano playing. For example, previous work has attempted to minimise for the variation of motor skill because the method of input is unrelated to the domain knowledge being measured (e.g. differences in typing ability would interfere with the signal when testing for programming knowledge [1]). For piano playing, however, more properties of the chunking signal may be able to be utilised, as it is directly related to the target skill - i.e. playing the piano. The transferral of chunking-based proficiency evaluation to this domain may be therefore significant because it suggests that the proposed method is robust enough to accommodate for, and even take advantage of, large variations in task-dependent motor skill between experts and novices.

In a similar vein, future research can investigate how the quality of proficiency measurement can depend on the alignment between an assessment method and the target of assessment. For example, to gain chunking metrics, the proposed assessment requires participants to complete a memory task that they would likely not encounter during normal piano playing scenarios: producing a long, unseen chord sequence from memory without a defined tempo. It is important that assessment tasks are designed in such a way that they produce measurements of the appropriate skills pertaining to the target learning objective, and not some kind of arbitrary skill that is too abstract and thus irrelevant to general piano playing.

6 CONCLUSION

We envision digital learning environments that operate in what appears almost like a niche in today's AI landscape, where hyper-automation is often the most obvious - but not necessarily most useful - approach: Systems that support humans in performing their tasks in a collaborative fashion rather than systems that are trying to replace them entirely. By using chunking as a framework for informing the mechanics of digital learning in creative domains, the proposed method of digital assessment could employ behavioural metrics to model piano players' skill based on the richness of associations that they have with musical material, as opposed to the accuracy with which they can play specific material. We hope that this will be useful to augment piano tuition and aid creativity and skill in a way that does not require algorithms to make overly rigid value judgements in what is largely an aesthetic domain, such as determining what may or may not be a "good" jazz improvisation, for example. Beyond this, we believe that chunking metrics could be useful for informing AI models that could drive the selection and presentation of learning content in assistive technologies, to more flexibly adapt to student proficiency, and with the aim to yield more effective tuition and support long-term motivation. While these systems are currently a long way off, the results of this study have shown promising correlations between the proposed metrics and observed proficiencies, and clearly motivate further research in this area.

ACKNOWLEDGEMENTS

We thank the reviewers of this paper for their useful feedback. We also thank Dr. Simon Bowes and Prof. Peter Cheng for the insightful discussions on embodied perspectives and memory chunking, and the latter also for providing additional comments on the final draft.

REFERENCES

- [1] Noorah Albehaijan and Peter C.-H. Cheng, 'Measuring Programming Competence by Assessing Chunk Structures in a Code Transcription Task', *Proceedings of the Annual Meeting of the Cognitive Science Society*, **41**(0), (2019).
- [2] John R. Anderson, *Learning and memory: an integrated approach*, Wiley, New York, N.Y., 2nd ed. edn., 2000.
- [3] Dana H. Ballard, Mary M. Hayhoe, Polly K. Pook, and Rajesh P. N. Rao, 'Deictic codes for the embodiment of cognition', *Behavioral and Brain Sciences*, **20**(4), 723–742, (December 1997).
- [4] William G. Chase and Herbert A. Simon, 'THE MIND'S EYE IN CHESS', in *Visual Information Processing*, ed., WILLIAM G. CHASE, 215–281, Academic Press, (1973).
- [5] Peter C.-H. Cheng, 'Analyzing chunk pauses to measure mathematical competence: Copying equations using 'centre-click' interaction', *Proceedings of the Annual Meeting of the Cognitive Science Society*, **37**(0), (2015).
- [6] Peter CH Cheng and Hector Rojas-Anaya, 'Measuring mathematic formula writing competence: An application of graphical protocol analysis', in *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, (2007).
- [7] Rachel Wooryung Dueck, *Developing a Progressive Chunking Technique in Music Literacy: An Analysis of the Method Books and Sight-Reading Materials for Beginner Piano Students*, Ph.D. dissertation, Université d'Ottawa / University of Ottawa, November 2023.
- [8] David Dunning, 'Chapter five - The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance', volume 44 of *Advances in Experimental Social Psychology*, 247–296, Academic Press, (2011).
- [9] Dennis E Egan and Barry J Schwartz, 'Chunking in recall of symbolic drawings', *Memory & cognition*, **7**(2), 149–158, (1979).
- [10] Fernand Gobet, Peter C R Lane, Steve Croker, Peter C-H Cheng, Gary Jones, Iain Oliver, and Julian M Pine, 'Chunking mechanisms in human learning', Technical report, (2001). Publication Title: TRENDS in Cognitive Sciences Volume: 5 Issue: 6.
- [11] Hadeel Bakr M. Ismail, *Competence assessment by stimulus matching (CASM): a novel approach to language assessment by chunk transcription*, thesis, University of Sussex, March 2025.
- [12] Lucas Lörch, Expertise in processing musical notation. An investigation of chunking, working memory, and eye movements, 2023. Place: Mannheim.
- [13] Richard E. Mayer, *Multimedia Learning*, Cambridge University Press, July 2020.
- [14] Robert S McLean and Lee W Gregg, 'Effects of induced chunking on temporal aspects of serial recitation.', *Journal of experimental psychology*, **74**(4p1), 455, (1967).
- [15] Timothy P. McNamara, James K. Hardy, and Stephen C. Hirtle, 'Subjective hierarchies in spatial memory', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**(2), 211–227, (1989).
- [16] George A. Miller, 'The magical number seven, plus or minus two: Some limits on our capacity for processing information', *Psychological Review*, **63**(2), 81–97, (1956). Place: US.
- [17] Unity Technologies. Unity game engine, 2026. <https://unity.com>, Accessed: 2026-08-06.
- [18] Margaret Wilson, 'Six views of embodied cognition', *Psychonomic Bulletin & Review*, **9**(4), 625–636, (December 2002).

Confusion, Agency, and Metacognitive Instruction in an Intelligent Tutoring System: An Ability-Moderated Account

Dmitri Droujkov and Sutapa Dey Tithi and Tiffany Barnes¹

Abstract. This study investigates the interaction between confusion, frustration, and agency under metacognitive instruction in a logic proof tutoring system ($N = 249$). The tutor teaches two domain-specific metacognitive skills, a Plan/Act/Reflect cycle adapted for logic proof construction and a Target Variable Strategy (TVS) for backward-chaining through logical rules, under varying degrees of learner agency (forced, nudge, optional). Using Control-Value Theory (CVT) as the explanatory framework, we report three findings. First, prior ability moderates which affect responds to metacognitive instruction. High-ability students show reduced frustration and increased confusion, while low-ability students show no affect differences. We call this a frustration-confusion swap: metacognitive instruction reduces task-level struggle while introducing metacognitive-level challenge. Second, nudge prompting generates more confusion than forced prompting in high-ability students. Prior work showed that agency moderates the functional consequences of confusion. Our data extend this by showing that agency also moderates confusion generation. Third, problem-type choice, an independent agency dimension, trends in the same direction. The results suggest that confusion under metacognitive instruction reflects effortful learning of domain-specific metacognitive skills. Furthermore, agency moderates both the generation and functional consequences of confusion.

1 INTRODUCTION

Intelligent tutoring systems (ITSs) increasingly embed metacognitive support alongside domain instruction [2, 24, 6], yet relatively little is known about how students *experience* this support affectively. ITS Research on affect has focused primarily on emotions during domain problem solving [9, 4]. However, even when affective states are measured alongside metacognitive scaffolding [5, 11, 3], prior research has treated affect as related to learning rather than as a response to the metacognitive instruction. Even less is known about how student affective experience varies by prior ability or the degree of agency afforded by the system. This paper investigates the relationship between confusion, agency, and metacognitive instruction in a logic proof tutoring system, using Control-Value Theory (CVT) as the explanatory framework.

1.1 Metacognitive instruction vs. scaffolding

Metacognitive support in ITSs can take two forms. *Scaffolding* provides temporary assistance for skills students are developing, fading

as competence grows [28, 19]. *Instruction* teaches new strategies intended for permanent internalization, constituting a genuine learning experience rather than temporary support [26].

The Deep Thought (DT) ITS used in this study teaches two domain-specific metacognitive skills. First, it teaches the Target Variable Strategy (TVS), a backward-chaining strategy in which students identify the target conclusion, determine which variables to derive, and select appropriate logical rules. Although students encountered a version of TVS adapted for probability earlier in the same course [1], the propositional logic-specific TVS requires different variable identification and rule selection strategies, managing forward and backward chaining through 16 logical rules rather than probability calculations. This makes TVS in logic a substantively distinct skill. Second, DT provides instruction in a Plan/Act/Reflect cycle adapted for logic proof construction, in the tradition of [22]. In the Plan phase, students enter goals to derive into a plan stack, specifying parent lines, elimination terms, and rules. The plan interface supports both forward chaining (from available logic statements called premises) and backward chaining (from the target conclusion). During planning, TVS is one backward-chaining strategy students may apply but are not required to. In the Act phase, students populate any remaining parameters for each plan step and execute it by clicking a “Do it” button that applies the planned rule on the proof canvas. Plan and Act may be interleaved step by step rather than completed as a full plan before any execution. In the Reflect phase, an end-of-problem dialog prompts students to type a free-text reflection on how the planning strategy helped them construct the proof, and to report mental effort and help preferences on Likert scales. A post-tutor reflection dialog at the session end assesses perceived mastery of the strategies. Neither Plan/Act/Reflect nor the logic-adapted TVS is a scaffold for an existing skill; both are genuinely new instruction.

The instructional goal is internalization, not fading. Unlike scaffolding that is withdrawn as competence develops, Plan/Act/Reflect and TVS are strategies students are expected to adopt as permanent components of their problem-solving repertoire. The forced, nudge, and optional prompting modes vary how the tutor promotes initial engagement with these strategies. This distinction matters for interpreting affective responses. If metacognitive support were purely scaffolding, confusion might signal that the support is unnecessary, reflecting a mismatch between the student’s existing competence and the imposed structure [13]. But if the support is genuine instruction in new metacognitive skills, confusion may instead signal productive cognitive disequilibrium from acquiring unfamiliar competencies [10, 14, 27]. Domain ability does not guarantee metacognitive proficiency [26]. A student who solves logic problems effectively

¹ North Carolina State University, Raleigh, NC, USA, ddroujk@ncsu.edu

may still have no experience with structured planning or backward-chaining strategies adapted for formal proofs.

1.2 Control-Value Theory

Pekrun’s [20] Control-Value Theory (CVT) proposes that learners’ appraisals of control and value shape achievement emotions and their consequences for performance, motivation, and well-being. In the original CVT framework, frustration is classified as an activity-related achievement emotion, experienced when a student expects failure and anticipates negative consequences, oriented toward the task itself rather than the learning process. The updated CVT [21] expanded its scope to a general theory of human emotion and introduced a three-dimensional taxonomy. CVT classifies confusion as an epistemic emotion appraised in response to incongruity that could be modeled as:

$$\text{Confusion} \propto \frac{\text{Incongruity} \cdot \text{Value}}{\text{Control}} \quad (1)$$

Confusion increases with the degree of incongruity encountered and the value assigned to the task, and decreases with perceived control. Agency, the degree to which a student has choice over their own behavior, maps onto perceived control in this formulation. Metacognitive support that reduces agency lowers perceived control and should therefore increase confusion when learners encounter incongruity.

1.3 Frustration and confusion as distinct constructs

Frustration and confusion are frequently treated as interchangeable negative affects in tutoring system research, yet they have distinct antecedents, durations, and learning consequences. D’Mello and Graesser [9] demonstrated that confusion arises at productive impasses, moments of cognitive disequilibrium that prompt deeper processing, while frustration signals unresolved impasse and predicts disengagement. D’Mello et al. [10] further showed that confusion predicts learning gains under conditions of productive cognitive engagement, while frustration does not. They also demonstrated that prior knowledge moderates confusion’s effects: “it takes a modicum of knowledge to know what one knows and does not know,” so low-knowledge learners may fail to detect impasses that would trigger confusion, while high-knowledge learners can both detect and resolve the resulting disequilibrium. Baker et al. [4] extended this analysis, showing that confusion and frustration occupy distinct but related positions on a shared affective continuum with different antecedents and consequences.

CVT specifies the appraisal mechanism that produces confusion (high incongruity and low control), while D’Mello and Graesser specify the functional outcome (deeper processing, if the confusion is resolved). Together, they predict that the *source* of incongruity determines whether confusion is productive: incongruity from encountering genuinely new material (metacognitive instruction) should produce productive confusion, while incongruity from unnecessary constraint (scaffolding mismatch) should not.

1.4 Two sources of incongruity

When students receive metacognitive support alongside domain instruction, the default interpretation of any resulting confusion is

scaffolding mismatch: students with established problem-solving approaches encounter imposed structure that conflicts with their existing strategies. On this account, the structure itself is the source of incongruity, and the resulting confusion reflects unnecessary cognitive load rather than productive learning [13].

However, when the metacognitive support constitutes genuine instruction in new skills rather than scaffolding of existing ones, an alternative interpretation applies. Plan/Act/Reflect and TVS are domain-specific strategies intended for internalization; neither is a scaffold for a skill students already possess. On this account, incongruity arises from acquiring unfamiliar metacognitive competencies, and confusion reflects *effortful metacognitive learning*, productive cognitive disequilibrium from reconciling new strategies with existing practice.

These sources are not mutually exclusive, but data can be used to distinguish between them. If confusion reflects pure scaffolding mismatch, then imposing metacognitive structure on students who already perform well should increase *both* frustration and confusion (or leave frustration unchanged while increasing confusion). If confusion instead reflects effortful metacognitive learning, we would expect frustration to *decrease* because the instruction reduces task-level struggle, while confusion *increases* from the new metacognitive challenge. We test this prediction in RQ1.

1.5 Agency and the confusion appraisal pathway

Self-determination theory (SDT; [25]) conceptualizes autonomy as a continuum of regulatory styles. The experimental conditions in this study vary student agency along a prompting mode dimension. *Forced* prompting drastically reduces choice over metacognitive behavior, requiring a completed plan on designated training problems before the student can proceed. *Nudge* reduces agency by reminding students to plan while preserving the option to decline. Finally, *Optional* allows planning mode but does not prompt for planning before starting the proof. SDT predicts that forced prompting produces introjected regulation (“I have to do this”), while nudge allows identified regulation (“I see this is useful, and I choose to do it”).

The CVT formula predicts that reducing agency (control) should increase confusion. However, agency may also moderate the *functional consequences* of confusion. Droujkov et al. [8] found that in a game-based learning environment, the inverse relationship between average confusion episode duration and intrinsic motivation appeared only under full agency. Under conditions that limited agency, this relationship disappeared. Their finding suggests that agency changes what confusion *does*, i.e., whether it connects to motivational outcomes, rather than simply whether it occurs. This raises the question of whether agency also moderates how confusion is *generated*: under forced prompting, students may comply behaviorally without the depth of cognitive engagement that produces confusion from incongruity.

Schoenfeld [26] demonstrated that expert problem solvers employ a plan-act-reflect cycle that novices lack, and that teaching this cycle explicitly improves mathematical problem-solving. However, the metacognitive control process itself is cognitively demanding, requiring sustained effortful self-regulation that students may experience as confusion even when it is productive. The degree to which students engage with this process, rather than merely performing it under compulsion, should determine whether confusion appears.

1.6 Prior work on confusion, ability, and metacognitive support in learning systems

Several lines of prior research converge on the questions addressed here. Lehman, D’Mello, and Graesser [15] directly tested who benefits from induced confusion, using false feedback to trigger cognitive disequilibrium. Only students with *both* high cognitive ability and high cognitive drive reported more confusion after false feedback. These students also showed transfer learning gains, but only when meaningfully confused. Students with low ability or low drive neither experienced productive confusion nor benefited from it. This finding is the closest precedent for our ability-moderated account: it establishes that confusion is not uniformly experienced and that individual differences determine whether confusion connects to learning.

Lodge and Kennedy [16] proposed a theoretical model of “zones of optimal and sub-optimal confusion,” integrating cognitive disequilibrium, desirable difficulties, productive failure, and impasse-driven learning. They argued that confusion is productive only when learners receive appropriate support at the right time and have the capacity to resolve the disequilibrium. Our two-source incongruity framework (Section 1.4) aligns with their model: scaffolding mismatch places learners in a zone of sub-optimal confusion, while effortful metacognitive learning places them in the optimal zone.

Richey et al. [23] examined confusion and frustration jointly, measured as “confrustion,” in an erroneous examples tutoring system. Confrustion was negatively correlated with both immediate and delayed posttest performance across conditions, yet students in the erroneous-example condition experienced more confrustion than problem-solving controls and still outperformed them on the delayed posttest — the condition-level benefit emerged despite, not because of, the elevated confrustion. Notably, Richey et al. treated confusion and frustration as a combined construct. Our data suggest they respond to different antecedents and should be measured separately.

Azevedo et al. [3] studied MetaTutor, an ITS that uses pedagogical agents to scaffold self-regulated learning. More intense initial prompting induced higher confusion in *low* prior knowledge students. MetaTutor’s prompts scaffold existing SRL processes, while DeepThought’s Plan/Act/Reflect and TVS are domain-specific strategies intended for internalization, not temporary support for skills students already possess. The direction of the ability-confusion interaction may depend on whether the metacognitive support is scaffolding (confusion in those who lack the skill) or instruction (confusion in those who must reconcile new strategies with existing competence).

1.7 Research questions

Three questions test the CVT account, each addressing a different term in Equation 1:

- RQ1** How does prior ability moderate which affect, frustration or confusion, responds to metacognitive instruction? (Tests *incongruity*: ability determines whether dominant incongruity is task-level or metacognitive-level.)
- RQ2** How does the degree of agency preserved by prompting mode (forced vs. nudge) modulate confusion in high-ability students? (Tests *control*: does agency moderate not only the functional consequences of confusion but also its generation?)
- RQ3** How does problem-type choice, an independent agency dimension, further modulate confusion? (Tests a second, independent manipulation of control.)

We additionally test the planning $\times$ pretest interaction on confusion with pretest as a continuous moderator.

2 METHOD

2.1 Participants

Participants were 249 undergraduate Computer Science students in a discrete mathematics course at a large public university. The study was a required lab assignment (Spring 2026). Students were randomly assigned to one of seven experimental conditions ($n = 34$ to 37 per condition).

2.2 Tutoring system

DeepThought is an intelligent tutoring system for teaching propositional logic proof construction [18, 17]. The interface presents the given premises as labeled nodes on a canvas-based workspace and the target conclusion at the bottom. Students iteratively select one or more premise nodes, choose an appropriate logic rule, and apply it to generate a new derived node until they complete the proof. The tutor provides immediate feedback on rule applications and adaptive problem sequencing. The tutor supports forward, backward, and indirect proof construction strategies.

The metacognitive component comprises the two skills: a logic-specific TVS (backward-chaining) and a Plan/Act/Reflect cycle in which students build a plan stack (forward or backward, TVS optional) and execute steps via a “Do it” button, interleaving planning and execution. After each problem, a reflection dialog collects a free-text reflection plus mental-effort (9-point) and help-preference ratings.

Table 1. Study procedure by level.

Level	Problems	Content	Types
1	5	Pretest & introduction	MWE, WE_A, PS, RM
P	3	Planning instruction	WE-P1, WE-P2, PS-PDF
2	4	Plan training begins	PS, PS-PDF, PPE, Buggy
3	4	Plan training continues	PS, PS-PDF, PPE, Buggy
4–6	4 each	Adaptive practice	PS, PPE, Buggy
7	6	Posttest	PS only

2.3 Procedure

The lab averaged approximately 5 hours across 8 levels with 3 to 6 problems each (Table 1).

Level 1 introduced the proof interface through mixed worked examples (MWE), guided worked examples (WE_A), problem-solving (PS), and a rule mastery (RM) exercise. Level P, inserted between Levels 1 and 2, provided planning instruction through two planning worked examples (WE-P1 focuses on plan creation via TVS; WE-P2 focuses on plan execution and parameter filling) followed by a planning dialog problem (PS-PDF), in which the tutor prompts the student through each step of TVS-based planning and then has them execute their own plan. Beginning in Levels 2 and 3, metacognitive conditions diverged: students in planning conditions applied Plan/Act/Reflect and TVS according to their assigned prompting mode. Levels 4 through 6 provided extended practice with adaptive problem sequencing (random or DRL-based, depending on condition). Level 7 served as the posttest, six novel problems solved without metacognitive prompting.

Three problem-type formats appeared from Level 2 onward. *PS* (problem-solving, the primary assessment format) requires the student to construct the entire proof from the given premises to the conclusion. *PPE* (partially worked problem example) presents a subset of the proof’s intermediate nodes already on the workspace, and the student fills in the missing derivation steps. *Buggy* presents an erroneous worked example, and the student identifies and corrects faulty steps. The posttest (Level 7) used PS exclusively to assess unprompted performance.

2.4 Experimental conditions

Seven conditions varied along three dimensions (Table 2).

Table 2. Experimental conditions.

tc	Prompting	Selection	Choice	Agency
1	Optional	Random	No	Highest
2	Forced	Random	No	Lowest
3	Nudge	Random	No	Medium
4	Forced	DRL	No	Lowest
5	Nudge	DRL	No	Medium
6	Forced	DRL	Yes	Low
7	Nudge	DRL	Yes	Higher

Prompting mode varied the degree of agency over metacognitive behavior. In *forced* conditions (tc = 2, 4, 6), four designated training problems (2.4, 2.5, 3.4, 3.5) were delivered as guided planning dialogs (PS-PDF) that required a completed plan before the student could proceed; on the remaining problems, planning tools were available but not mandatory. In *nudge* conditions (tc = 3, 5, 7), students received a reminder to plan but could decline and proceed without planning. In the *optional* condition (tc = 1), no planning prompts were presented. Students could use planning tools voluntarily.

Selection strategy determined problem sequencing. In random conditions (tc = 1 to 3), problems within each level were drawn uniformly at random from the level’s pool of allowed formats (PS, PPE, Buggy). In DRL conditions (tc = 4 to 7), a deep reinforcement learning policy selected the next problem type based on a 72-feature student model updated after each problem, capturing per-step performance, rule usage, time-on-task, and recent error patterns. From the student’s perspective, the two strategies differ in within-level adaptivity, not in the set of formats available.

Problem-type choice was nested within DRL conditions. In choice conditions (tc = 6, 7), before each new problem in Levels 4 through 6, the tutor displayed a selection dialog listing three options (PS, PPE, Buggy); the student chose the format of the upcoming problem. The DRL policy still controlled *which* specific problem instance was delivered for the chosen format, so the student selected the format but not the content. In all other conditions, the DRL policy or random assignment determined the problem type. Problem-type choice was thus an agency dimension orthogonal to prompting mode: a student in a forced prompting condition could still hold problem-type choice, and a student in a nudge condition could still lack it.

2.5 Measures

Prior ability. Pretest performance indexes *domain ability in propositional logic proof construction*, not metacognitive regulation or self-reported confidence. It was computed as the mean problem-Score across two Level 1 problem-solving (PS) problems (1.3 and

1.4), administered before any planning instruction (Level P), so the measure is uncontaminated by exposure to TVS or Plan/Act/Reflect. The problemScore composite equally weights rule accuracy, problem completion time, and solution length, normalized by interquartile range within each problem. Pretest scores ranged from 0.31 to 0.93 ($M = 0.693$, $SD = 0.147$, $Mdn = 0.722$). Consistent with the distinction drawn in Section 1.1, high domain ability on this measure does not entail metacognitive proficiency [26].

Posttest. Mean problemScore across six Level 7 problems (7.1 to 7.6), all solved without metacognitive prompting. Problems 7.1 and 7.2 were isomorphic to pretest; 7.3 to 7.6 required transfer to novel proof structures ($M = 0.704$, $SD = 0.116$, $n = 247$).

Affect. Students self-reported their affective state by clicking one of five emoji buttons available throughout the session: Stuck/Frustrated, Confused, Bored, Flow, or Delighted. Reports were voluntary and could be submitted at any time. A total of 7,722 reports were collected from 174 of 249 students (70%), with a mean of 44.4 reports per reporting student ($Mdn = 20.0$). Primary measures were per-student proportions: frustrated_prop (59.5% of all reports) and confused_prop (20.4%). Students with no affect reports were excluded from affect analyses.

Plan quality. For problems with planning, plan quality ($M = 0.875$, $SD = 0.087$) was the ratio of correct plan steps to total steps: $\text{planCorrectCount} / (\text{planCorrectCount} + \text{planErrorCount})$.

Voluntary metacognitive behavior: Measured on PS problems only (where planning was not mandatory regardless of condition): voluntary planning rate (proportion of PS problems with any plan goal selection; $M = 0.046$, $SD = 0.077$) and TVS adoption rate (proportion of PS problems with any mode toggle to term view; $M = 0.066$, $SD = 0.084$).

2.6 Analysis

Three planned contrasts addressed the research questions: (1) Planning (tc = 2 to 7) vs. Optional (tc = 1), testing whether metacognitive instruction produces different affective profiles; (2) Forced vs. Nudge (tc = 2, 4, 6 vs. tc = 3, 5, 7), testing whether preserved agency modulates confusion; (3) Choice vs. No Choice (tc = 6, 7 vs. tc = 4, 5, within DRL), testing an independent agency dimension. Each contrast was tested with Welch’s *t*-test and Mann-Whitney *U*, with Cohen’s *d* and 95% CIs. Kruskal-Wallis tests across all seven conditions preceded planned contrasts.

Ability moderation. Each contrast was tested separately within bottom 50% and top 50%, defined by a median split on pretest ($Mdn = 0.722$): $n = 125$ (bottom) and $n = 124$ (top) on the affect sample ($N = 249$, used for Findings 1–3), and $n = 125$ and $n = 122$ on the posttest-complete sample ($n = 247$). To confirm that ability moderation was not a dichotomization artifact, we fit OLS regression with centered pretest as a continuous moderator and tested the planning $\times$ pretest interaction on confused_prop and frustrated_prop.

Exploratory analyses. Within the top 50%, we compared forced and nudge on plan quality, voluntary planning rate, and TVS adoption rate to test whether the same conditions producing more confusion also produced better metacognitive engagement (convergent validity for the effortful learning interpretation). We correlated confused_prop and frustrated_prop with posttest scores, voluntary planning rate, TVS adoption rate, and plan quality within ability groups to test whether confusion was associated with immediate learning outcomes. The ability-split analyses are post-hoc and exploratory. We report them as pattern-generating evidence warranting pre-registered

replication, not confirmatory tests.

3 RESULTS

3.1 Full-sample affect: no condition differences

In the full sample ($N = 174$ with affect data), no significant differences in frustrated or confused proportions were observed across the seven conditions (Kruskal-Wallis: frustrated $H = 6.19$, $p = .40$; confused $H = 6.09$, $p = .41$). Frustrated/stuck was the dominant reported affect across all conditions ($M = 0.47$ to 0.62 , 59.5% of all reports), while confusion was less frequent ($M = 0.10$ to 0.28 , 20.4%). This null provides the baseline against which the ability-moderated findings below should be interpreted.

3.2 Finding 1: ability moderates which affect responds (RQ1)

The full-sample null masks a crossover by prior ability (Table 3).

Table 3. Affect by planning condition and ability group.

Affect	Contrast	Bottom 50%	Top 50%
Frustrated	Plan vs Opt	$d = 0.08$ $p = .79$	$d = -0.68^*$ $p = .043$ CI $[-1.33, -0.04]$
Confused	Plan vs Opt	$d = -0.55$ $p = .15$	$d = 0.60$ $p = .24$ (MW) CI $[-0.04, 1.24]$

For top 50% students, metacognitive instruction significantly *reduced* frustration (Planning $M = 0.432$, Optional $M = 0.714$; $d = -0.681$, $p = .043$) and *increased* confusion (Planning $M = 0.217$, Optional $M = 0.029$). The confusion increase is not a shift in the typical student (both medians = 0): the parametric $d = 0.601$, $p < .001$ is inflated by near-zero variance in the small Optional cell ($n = 11$), and the rank test is non-significant (Mann-Whitney $p = .24$). It instead reflects a higher report-level rate of confusion, where high-ability planning students were more likely than Optional students to express confusion on any given report (mixed/GEE logistic OR = 13.4, 95% CI $[2.1, 87.8]$, $p = .007$). This frustration-confusion swap, a robust frustration reduction alongside an elevated confusion rate, is the central finding. Instruction in Plan/Act/Reflect and TVS reduced task-level struggle while introducing metacognitive-level challenge. For bottom 50% students, neither affect differed significantly by condition. Frustration was high across the board and confusion was low, consistent with task-level incongruity dominating the affective experience regardless of metacognitive instruction.

Continuous moderation. Regressing confused_prop on planning condition, centered pretest, and their interaction yielded a significant interaction ($\beta = 1.353$, $t = 2.98$, $p = .003$; $R^2 = .052$ vs. .003 for main effects; $F = 8.89$, $p = .003$). The effect of planning instruction on confusion increases continuously with prior ability. This interaction is robust to estimator (HC3 robust SE $p = .003$; fractional-logit $p = .002$) and does not depend on the median split; we treat it as the primary test for RQ1, with the dichotomized contrast in Table 3 as illustration. A parallel model for frustrated_prop showed a directionally opposite interaction ($\beta = -0.980$, $t = -1.67$, $p = .096$).

Swap by prompting mode. Decomposing the swap by mode shows the two arms behave differently. Relative to Optional, frustration fell under both prompting modes (nudge $d = -0.79$, $p = .02$; forced $d = -0.57$, $p = .09$), and the two modes did not differ from each other on frustration (forced vs. nudge $d = 0.27$, $p = .27$). The confusion increase, by contrast, was concentrated in nudge: forced was indistinguishable from Optional on confusion (Mann-Whitney $p = .71$), and the forced vs. nudge difference was significant (Finding 2). The frustration reduction is thus a general effect of metacognitive instruction, whereas the agency grading that reaches significance is specific to confusion.

3.3 Finding 2: nudge > forced on confusion (RQ2)

Within planning conditions, the forced vs. nudge prompting contrast was tested separately by ability group for both confused and frustrated proportions (Table 4).

Table 4. Confused and frustrated proportions by prompting mode and ability group.

Affect	Ability	Forced M (SD)	Nudge M (SD)	d	p	95% CI
Confused	Top 50%	0.129 (0.250)	0.300 (0.383)	-0.524	.030	$[-1.00, -0.05]$
Confused	Bottom 50%	0.223 (0.308)	0.189 (0.312)	0.109	.63	$[-0.33, 0.55]$
Frustrated	Top 50%	0.489 (0.395)	0.378 (0.436)	0.266	.27	$[-0.21, 0.74]$
Frustrated	Bottom 50%	0.523 (0.424)	0.638 (0.367)	-0.290	.20	$[-0.73, 0.15]$

Among top 50% students, nudge produced significantly more confusion than forced ($d = -0.524$, $p = .030$). The 95% CI excludes zero. Among bottom 50%, forced and nudge did not differ on confusion ($d = 0.109$, $p = .63$). Because the forced and nudge groups each include a choice condition (tc6 forced, tc7 nudge), this contrast is not independent of problem-type choice; restricted to the no-choice conditions (tc2–5) the effect attenuates but stays directionally consistent ($d = -0.49$, Welch $p = .10$, Mann-Whitney $p = .04$).

Frustration did not reach significance in either ability group, but the directional pattern is informative. In top 50% students, nudge produced numerically *less* frustration than forced ($d = 0.266$, $p = .27$), the directional opposite of the confusion effect: the same prompting mode that elevated confusion also tended to suppress frustration. In bottom 50% students, the direction reversed: nudge produced numerically *more* frustration than forced ($d = -0.290$, $p = .20$), with confusion essentially unchanged. The contrast between ability groups argues against an interpretation that nudge simply increases all negative affect. For high-ability students, preserved agency converts task-level frustration into metacognitive-level confusion. For lower-ability students, who lack the domain resources to capitalize on the freedom, the same preserved agency tends to amplify task-level frustration without engaging the metacognitive challenge that would generate confusion.

This pattern is counterintuitive under a naive reading of the CVT formula: nudge preserves more agency (higher control), which should predict *less* confusion. Droujkov et al. [8] showed that agency moderates the functional relationship between confusion episode dynamics and motivation, where the inverse relationship between average confusion episode duration and intrinsic motivation appeared only under full agency. Our finding extends this: agency moderates not only what confusion *does* but whether it is *generated*. Forced prompting may reduce agency so completely that students comply procedurally without the depth of cognitive engagement required for incongruity to register as confusion. Nudge, by preserving the option

to decline, allows identified regulation [25]. Students who choose to engage with the strategy process deeply enough for the incongruity to produce confusion rather than mere procedural compliance.

Supporting evidence from metacognitive behavior. Within the top 50%, nudge produced marginally higher plan quality ($d = -0.358$, $p = .068$), while voluntary planning rate ($d = 0.276$, $p = .16$) and TVS adoption ($d = 0.064$, $p = .74$) did not differ. In the bottom 50%, the plan quality advantage of nudge over forced was large ($d = -0.993$, $p < .0001$). Because forced prompting mandated a completed plan on the designated planning-dialog problems while nudge did not, forced students produced more plan steps on those problems; the nudge ratio advantage therefore reflects more selective, higher-accuracy engagement among those who choose to plan rather than a greater volume of correct metacognitive practice.

3.4 Finding 3: choice trends toward more confusion (RQ3)

A second, independent agency dimension, problem-type choice ($tc = 6, 7$ vs. $tc = 4, 5$, within DRL), showed a parallel pattern in top 50% students (Table 5).

Table 5. Affect by choice availability (top 50% only).

Affect	Choice M (SD)	No Choice M (SD)	d	p	95% CI
Confused	0.327 (0.391)	0.151 (0.289)	0.524	.099	[-0.06, 1.11]
Frustrated	0.440 (0.427)	0.437 (0.421)	0.005	.99	[-0.57, 0.58]

The effect on confusion was in the same direction and of the same magnitude as Finding 2 ($d = 0.524$), but marginal ($p = .099$). The 95% CI includes zero. Frustration was entirely unaffected by choice ($d = 0.005$, $p = .99$). This result is consistent with the CVT account, where a second, independent manipulation of perceived control produces the same directional effect on confusion without affecting frustration, but it does not reach conventional significance with the available sample size ($n = 20$ choice, $n = 27$ no choice within the top 50%).

Choice conditions also reported significantly fewer total affect events ($M = 19.2$ vs. 33.4 , Mann-Whitney $p = .009$, $d = -0.33$), suggesting problem-type choice provides an agency outlet reducing the overall impulse to report affect. When students in choice conditions did report, a greater proportion of their reports were confusion rather than frustration.

3.5 Confusion and learning outcomes

Within the top 50%, confusion was unrelated to overall posttest (Spearman $\rho = -0.011$, $p = .92$, $n = 79$), isomorphic posttest ($\rho = -0.009$, $p = .94$), and transfer posttest ($\rho = -0.023$, $p = .84$). Frustration showed a weak negative association with isomorphic posttest ($\rho = -0.226$, $p = .045$) but not overall or transfer posttest. In the bottom 50%, neither affect correlated with any posttest measure (all $|\rho| < 0.17$, all $p > .11$).

Confusion was also unrelated to metacognitive behavior within the top 50%: confused_prop did not correlate with voluntary planning ($\rho = 0.145$, $p = .20$), TVS adoption ($\rho = 0.079$, $p = .49$), or plan quality ($\rho = 0.044$, $p = .70$). The confusion observed under metacognitive instruction is neither productive (associated with better learning) nor harmful (associated with worse learning) within a single session. It is orthogonal to immediate performance outcomes.

4 DISCUSSION

4.1 A unified CVT account

The three findings converge on a single mechanism within CVT's framework (Equation 1).

Finding 1 addresses the *incongruity* term. For high-ability students, Plan/Act/Reflect and TVS are genuinely new metacognitive skills that create strategy-level incongruity. The instructed approach extends beyond or conflicts with their existing problem-solving methods. These students have sufficient domain competence to handle the proof tasks themselves (frustration drops), but they experience confusion from the unfamiliar metacognitive process. For lower-ability students, the dominant incongruity is at the task level. The proofs themselves are challenging, producing frustration rather than confusion.

Finding 2 addresses the *control* term, but not in the direction a naive CVT reading predicts. Forced prompting reduces control more than nudge, yet produces *less* confusion. Droujkov et al. [8] showed that agency moderates the functional consequences of confusion, where the confusion-motivation link only appeared under full agency. Our data go further, showing that agency also moderates the *generation* of confusion. Forced prompting produces behavioral compliance without the cognitive engagement needed for incongruity to be appraised. Nudge preserves enough agency for identified regulation [25]. Students who choose to engage with the strategy process it deeply enough for the incongruity to register. Control operates not as a simple denominator but as a threshold: below a minimum level of agency, genuine engagement does not occur, and confusion is suppressed.

Finding 3 provides a partial replication with an independent manipulation of control. Problem-type choice, a second agency dimension, produces the same directional effect ($d = 0.524$) on confusion in high-ability students, though the effect is marginal. The convergence across two independent manipulations of agency, prompting mode (Finding 2) and problem-type choice (Finding 3), strengthens the inference that control is the operative CVT parameter, even though neither manipulation alone would be definitive.

4.2 Confusion as effortful metacognitive learning

The frustration-confusion swap in high-ability students (Finding 1) is the primary evidence favoring effortful learning over scaffolding mismatch. Three observations support this interpretation. First, frustration drops ($d = -0.681$) while confusion rises ($d = 0.601$). Pure mismatch predicts both negative affects increase or frustration remains flat. Second, nudge produces more confusion than forced ($d = -0.524$). If confusion were a reaction to unnecessary constraint, the more controlling condition should produce more. Third, nudge produces higher plan-quality ratios than forced (overall $d = 0.66$, bottom 50% $d = 0.993$; nudge higher in both). The conditions generating more confusion also generate higher-accuracy (though lower-volume) planning.

The two sources of incongruity, scaffolding mismatch and effortful learning, are not mutually exclusive, and the current data cannot definitively separate them. However, the overall pattern is more consistent with productive metacognitive challenge. D'Mello and Graesser [9] describe exactly this kind of productive confusion: cognitive disequilibrium at an impasse that, if resolved, drives deeper processing. Lodge and Kennedy's [16] model of optimal vs. sub-optimal confusion zones provides further theoretical grounding. The frustration-confusion swap suggests high-ability students are in the

optimal zone (productively challenged by new metacognitive skills), not the sub-optimal zone (overwhelmed by unnecessary constraint).

Plan/Act/Reflect and TVS create productive impasses because they are domain-specific skills requiring new learning, not generic “plan ahead” prompts (see Section 2.2). Even students who performed well on the domain tasks had no prior experience with these structured approaches to logic proof construction. As Schoenfeld [26] demonstrated, expert problem solvers employ metacognitive control processes that novices lack, and the control process itself is cognitively demanding even when the domain content is not.

The null confusion-posttest relationship (Section 3.5) neither supports nor undermines this interpretation. Richey et al. [23] observed a similar dissociation: confusion correlated negatively with posttest performance, yet the erroneous-example condition, despite more confusion than controls, outperformed them on the delayed posttest. Within a single 5-hour session, metacognitive skills may be too nascent to manifest as measurable performance gains. Students are still learning to plan and apply TVS, not yet reaping benefits from internalized strategies. This temporal dissociation between metacognitive learning and domain performance is expected when two skill dimensions are developing on different timescales.

4.3 Agency as moderator of confusion dynamics

Droujkov et al. [8] showed that agency moderates the *functional consequences* of confusion: the inverse relationship between the average duration of the confusion episode and intrinsic motivation appeared only under full agency. Our data extend this by showing that agency also moderates the *generation* of confusion in high-ability students ($d = -0.524$, $p = .030$ for prompting; $d = 0.524$, $p = .099$ for choice). Together, the two studies suggest that agency operates at multiple levels of the confusion process and that the relationship is ability-moderated: only high-ability students show the confusion response to agency manipulations, while lower-ability students’ affective responses are dominated by task-level challenge regardless of agency.

4.4 Design implications

These findings have four implications for the design of affect-aware tutoring systems.

First, confusion under metacognitive instruction is diagnostic rather than uniformly negative. In high-ability students, it signals engagement with new metacognitive skills, and removing instruction in response would terminate the productive cognitive disequilibrium the system was designed to support. Lodge and Kennedy’s [16] zones model suggests the appropriate response is to maintain the learner in this optimal confusion zone rather than eliminating confusion entirely.

Second, frustration and confusion have distinct antecedents and impacts and should be measured with sufficient resolution to recover that distinction [4], rather than collapsed into a single “confusion” construct [23]. Frustration in lower-ability students signals task-level struggle; persistence support, hints, or reduced problem difficulty are appropriate. Confusion in higher-ability students signals metacognitive-level challenge; continued engagement with the strategy may be sufficient.

Third, prompting should preserve agency. Nudge produces both higher confusion and higher-accuracy planning than forced prompting. The agency preserved by nudge enables the genuine cognitive engagement that drives both productive confusion and more accurate

planning among those who plan. Forced prompting achieves higher-volume but lower-accuracy planning on the mandated problems, with voluntary uptake comparable to nudge but less of the confusion that marks deeper processing.

Fourth, affect interpretation requires ability context, and the direction of the ability-confusion interaction depends on whether metacognitive support is scaffolding or instruction. Lehman et al. [15] found that only students with high cognitive ability and drive benefit from induced confusion, and Azevedo et al. [3] found that MetaTutor’s SRL prompts, which scaffold existing self-regulation processes, induced more confusion in low prior knowledge students, while our domain-specific metacognitive instruction induced more confusion in high-ability students. Systems should distinguish between scaffolding existing skills (where confusion signals insufficient support) and teaching new skills (where confusion signals productive engagement), and should not act on a detected affect without also conditioning on ability.

5 LIMITATIONS AND FUTURE DIRECTIONS

The affect measure relies on voluntary emoji self-reports, which depend on students’ willingness and ability to introspect on their affective states. Of 249 students, 174 (70%) reported any affect, and proportions are sensitive to reporting frequency. Behavioral or physiological measures of confusion would provide convergent validity.

The ability-moderation analyses are exploratory. The median split on pretest is post-hoc and not pre-registered. After splitting by both ability and condition, sample sizes reduce to $n = 15$ to 20, insufficient for confirmatory inference. The continuous moderation ($p = .003$) mitigates the dichotomization concern but does not address the post-hoc nature.

Finding 3 (choice effect) is marginal ($p = .099$) and should be interpreted cautiously. Problem-type choice is confounded with DRL-based problem selection. Choice was only available under DRL conditions ($tc = 6, 7$), so the choice effect cannot be fully isolated from the effects of the problem selection strategy. The forced vs. nudge contrast (Finding 2) is likewise not fully independent of choice, since the choice conditions sit within it; removing them attenuates the effect to $p = .10$ (Welch), though the nonparametric test remains significant ($p = .04$).

The CVT formula is used as a conceptual framework, not a quantitative model. The parameters (λ , incongruity, value, control) are not measured directly. We infer incongruity from the novelty of the metacognitive instruction, and control from the experimental manipulation of agency. Direct measurement of perceived control and perceived incongruity would allow more precise tests of the appraisal mechanism.

The effortful learning and scaffolding mismatch interpretations cannot be definitively distinguished. Think-aloud protocols or process-level confusion tracking could reveal whether confusion diminishes as metacognitive strategies become familiar (consistent with effortful learning) or persists throughout the session (more consistent with mismatch). Level-by-level affect data from this study show suggestive but underpowered trends: confusion in top 50% planning students peaks at training levels and decreases toward the posttest, but per-level cell sizes are too small for reliable inference.

Future work should pursue four directions. First, a pre-registered ability-stratified design with adequate per-cell power ($n = 30+$ per ability group $\times$ condition cell) to confirm the aptitude-treatment interaction on confusion [15]. Second, longitudinal tracking across multiple sessions, motivated by Richey et al.’s [23] finding that

confusion predicted delayed but not immediate posttest gains. Repeated sessions would test a central prediction of the productive-confusion hypothesis: confusion should diminish as students grow familiar with Plan/Act/Reflect and TVS, whereas confusion that persists across sessions would instead indicate a lasting scaffolding mismatch. Third, direct manipulation of CVT components to test the appraisal mechanism rather than inferring it from conditions. Fourth, a within-study comparison of scaffolding vs. instruction conditions to test whether the ability-confusion interaction reverses.

6 CONCLUSION

Confusion under metacognitive instruction in an intelligent tutoring system is not a uniform signal of difficulty. It is an ability-moderated, agency-dependent response that, in high-ability students, co-occurs with reduced frustration and more accurate planning when students choose to plan. Control-Value Theory provides a coherent framework: the source of incongruity (task-level vs. metacognitive-level) determines which affect responds, while the degree of agency moderates both the generation and the functional consequences of confusion. Although theoretical work has recognized confusion's productive potential [10, 16, 14], affect-aware system implementations typically treat detected confusion as a negative state to be remediated through hints or difficulty reduction [7, 12]. Our findings suggest this default response is inappropriate when confusion arises from metacognitive instruction in high-ability students, and that prompting designs preserving learner agency, nudge rather than forced, produce both richer affective engagement and better metacognitive outcomes.

REFERENCES

- [1] Mark Abdelshieed, Guojing Zhou, Mehak Maniktala, Tiffany Barnes, and Min Chi, 'Metacognition and motivation: The role of time-awareness in preparation for future learning', in *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42, (2020).
- [2] Roger Azevedo, Vincent Aleven, et al., *International handbook of metacognition and learning technologies*, volume 26, Springer, 2013.
- [3] Roger Azevedo, François Bouchet, Melissa Duffy, Jason Harley, Michelle Taub, Gregory Trevors, Elizabeth Cloude, Daryn Dever, Megan Wiedbusch, Franz Wortha, et al., 'Lessons learned and future directions of metatutor: Leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system', *Frontiers in Psychology*, **13**, 813632, (2022).
- [4] Ryan S Baker, Elizabeth Cloude, Juliana MAL Andres, and Zhanlan Wei, 'The frustration constellation: a new way of looking at confusion and frustration', *Cognitive Science*, **49**(1), e70035, (2025).
- [5] Annamarie Brosnihan, Megan Wiedbusch, Cameron Marano, Maral Karimi, Matthew Moreno, Tara Delgado, Oscar Tidwell, and Roger Azevedo, 'Balancing emotional and motivational self-regulatory processes during complex learning with intelligent tutoring systems', in *International Conference on Artificial Intelligence in Education*, pp. 205–217. Springer, (2025).
- [6] Daryn A Dever, Nathan A Sonnenfeld, Megan D Wiedbusch, S Grace Schmorrow, Mary Jean Amon, and Roger Azevedo, 'A complex systems approach to analyzing pedagogical agents' scaffolding of self-regulated learning within an intelligent tutoring system', *Metacognition and Learning*, **18**(3), 659–691, (2023).
- [7] Sidney D'Mello and Arthur Graesser, 'Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back', *ACM Transactions on Interactive Intelligent Systems (TiiS)*, **2**(4), 1–39, (2013).
- [8] Dmitri Droujkov, Andrew Emerson, Dan Carpenter, Xiaoyi Tian, Roger Azevedo, and Tiffany Barnes, 'Investigating the impact of confusion and agency on motivation in a game-based learning environment', in *International Conference on Artificial Intelligence in Education*, pp. 177–189. Springer, (2025).
- [9] Sidney D'Mello and Art Graesser, 'Dynamics of affective states during complex learning', *Learning and Instruction*, **22**(2), 145–157, (2012).
- [10] Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser, 'Confusion can be beneficial for learning', *Learning and Instruction*, **29**, 153–170, (2014).
- [11] Anastasia Efklides, 'Affect, epistemic emotions, metacognition, and self-regulated learning', *Teachers College Record*, **119**(13), 1–22, (2017).
- [12] Kate Forbes-Riley and Diane Litman, 'Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor', *Speech Communication*, **53**(9-10), 1115–1136, (2011).
- [13] Slava Kalyuga, 'Expertise reversal effect and its implications for learner-tailored instruction', *Educational psychology review*, **19**(4), 509–539, (2007).
- [14] Manu Kapur, 'Productive failure in learning math', *Cognitive science*, **38**(5), 1008–1022, (2014).
- [15] Blair Lehman, Sidney D'Mello, and Art Graesser, 'Who benefits from confusion induction during learning? an individual differences cluster analysis', in *International conference on artificial intelligence in education*, pp. 51–60. Springer, (2013).
- [16] Jason M Lodge, Gregor Kennedy, Lori Lockyer, Amael Arguel, and Mariya Pachman, 'Understanding difficulties and resulting confusion in learning: An integrative review', in *Frontiers in Education*, volume 3, p. 49. Frontiers Media SA, (2018).
- [17] Mehak Maniktala, Christa Cody, Tiffany Barnes, and Min Chi, 'Avoiding help avoidance: Using interface design changes to promote unsolicited hint usage in an intelligent tutor', *International Journal of Artificial Intelligence in Education*, **30**(4), 637–667, (2020).
- [18] Behrooz Mostafavi and Tiffany Barnes, 'Evolution of an intelligent deductive logic tutor using data-driven elements', *International Journal of Artificial Intelligence in Education*, **27**(1), 5–36, (2017).
- [19] Roy D Pea, 'The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity', in *Scaffolding*, 423–451, Psychology Press, (2018).
- [20] Reinhard Pekrun, 'The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice', *Educational psychology review*, **18**(4), 315–341, (2006).
- [21] Reinhard Pekrun, 'Control-value theory: From achievement emotion to a general theory of human emotions', *Educational Psychology Review*, **36**(3), 83, (2024).
- [22] G. Pólya, *How to Solve It*, Princeton University Press, 1945.
- [23] J Elizabeth Richey, Juan Miguel L Andres-Bray, Michael Mogessie, Richard Scruggs, Juliana MAL Andres, Jon R Star, Ryan S Baker, and Bruce M McLaren, 'More confusion and frustration, better learning: The impact of erroneous examples', *Computers & Education*, **139**, 173–190, (2019).
- [24] Ido Roll, Vincent Aleven, Bruce M McLaren, and Kenneth R Koedinger, 'Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system', *Learning and Instruction*, **21**(2), 267–280, (2011).
- [25] Richard M Ryan and Edward L Deci, 'Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being', *American psychologist*, **55**(1), 68, (2000).
- [26] Alan H Shoenfeld, 'Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics', *Handbook of research on mathematics teaching and learning*, (1992).
- [27] Kurt VanLehn, 'Toward a theory of impasse-driven learning', in *Learning issues for intelligent tutoring systems*, 19–41, Springer, (1988).
- [28] David Wood, Jerome S Bruner, and Gail Ross, 'The role of tutoring in problem solving', *Journal of child psychology and psychiatry*, **17**(2), 89–100, (1976).

Semantic Scaffolding: Augmenting Textual Structures with Domain-Specific Groupings for Accessible Data Exploration

Jonathan Zong¹²

Isabella Pedraza Pineros¹³

Mengzhu (Katie) Chen⁴

Daniel Hajas⁵

Arvind Satyanarayan⁶

Abstract. Drawing connections between interesting groupings of data and their real-world meaning is an important, yet difficult, part of encountering a new dataset. A lay reader might see an interesting visual pattern in a chart but lack the domain expertise to explain its meaning. Or, a reader might be familiar with a real-world concept but struggle to express it in terms of a dataset’s fields. In response, we developed *semantic scaffolding*, a technique for using domain-specific information from large language models (LLMs) to identify, explain, and formalize semantically meaningful data groupings. We present groupings in two ways: as *semantic bins*, which segment a field into domain-specific intervals and categories; and *data highlights*, which annotate subsets of data records with their real-world meaning. We demonstrate and evaluate this technique in Olli, an accessible visualization tool that exemplifies tensions around explicitly defining groupings while respecting the agency of readers to conduct independent data exploration. We conducted a study with 15 blind and low-vision (BLV) users and found that readers used semantic scaffolds to quickly understand the meaning of the data, but were often also critically aware of its influence on their interpretation.

1 Introduction

When a lay reader encounters a new dataset on a website or news article, an important part of their exploratory process is to identify interesting groupings of data, and explain them in terms of that data’s real-world meaning. Broadly, the real-world meaning of data is known as its *semantics*; domain-specific knowledge about data is how a reader determines whether a number “represent[s] a day of the month, or an age, or a measurement of height, or a unique code for a specific person, or a postal code for a neighborhood, or a position in space” [14] or other possible meanings. Effectively drawing connections between data and its semantics can be challenging for a lay reader. For instance, a reader might see an interesting visual pattern in a chart but lack the domain-specific knowledge to explain its meaning. In this case, difficulty arises because the reader does not know what they don’t know — it may be difficult to know how to begin to acquire the relevant context. Or, a reader might be familiar with a real-world concept but struggle to translate it into the terms of the dataset. For

example, a reader may not know how to approximate a subjective concept like “sports car” in terms of fields like `Horsepower` and `Miles.per.Gallon` — or possibly other fields in the dataset that were not in the initial visualization. These challenges can affect a reader’s comprehension of the data — as prior research has shown, a reader’s interpretation of a visualization is sensitive to differences in their prior knowledge and personal background [9, 16].

In response, we developed a technique called *semantic scaffolding* for using domain-specific information from large language models (LLMs) to identify, explain, and formalize semantically meaningful data groupings. Rather than use an LLM to generate summary descriptions or chat responses, we use it to return groupings as data structures that include a *name*, *explanation*, and *query predicate*. For instance, in the example cars dataset, an LLM might create a grouping named “Fuel Efficient Japanese Cars” with the following explanation: “This group represents cars from Japan that are known for their fuel efficiency, reflecting Japanese automotive engineering and consumer trends towards sustainable driving” (Figure 1). Crucially, the LLM associates this grouping with the following query predicate: `Miles.per.Gallon ≥ 25 ∧ Origin = Japan`. This allows a reader to connect the name and explanation with an explicitly-defined subset of data. We developed two types of interface elements for presenting semantic groupings to a reader: *semantic bins*, which segment a single field (i.e. column) into domain-specific intervals and categories; and *data highlights*, which annotate subsets of data records (i.e. rows) with their real world meaning. These two uses of semantic scaffolding in an interface serve distinct purposes. Semantic bins help a reader break down a single field into understandable pieces, facilitating navigation and exploration through a dataset. Data highlights help a reader quickly get an overview of a dataset, indicating potentially interesting subsets to begin exploring further.

We prototyped these designs via extensions to Olli [3], an accessible visualization system for screen reader users. Data accessibility is a rich context in which to evaluate our work because current tools for blind and low vision (BLV) screen reader users amplify the challenges lay readers of visualizations face. First, sighted readers often rely on their visual perception to identify interesting groupings in a visualization, but screen reader interfaces rarely afford a similar type of overview. Second, BLV readers highly value the agency to independently explore data, conduct an open-ended interpretive process, and arrive at their own conclusions about the data [13]. For instance, research has shown that screen reader users find descriptions less useful when they over-emphasize contextual and domain-specific infor-

¹ Co-first authors

² University of Colorado Boulder, USA, e-mail: jzong@colorado.edu

³ MIT CSAIL, USA, e-mail: ipedraza@mit.edu

⁴ MIT CSAIL, USA, e-mail: mzc219@mit.edu

⁵ UCL Computer Science, UK, e-mail: d.hajas@ucl.ac.uk

⁶ MIT CSAIL, USA, e-mail: arvindsatya@mit.edu

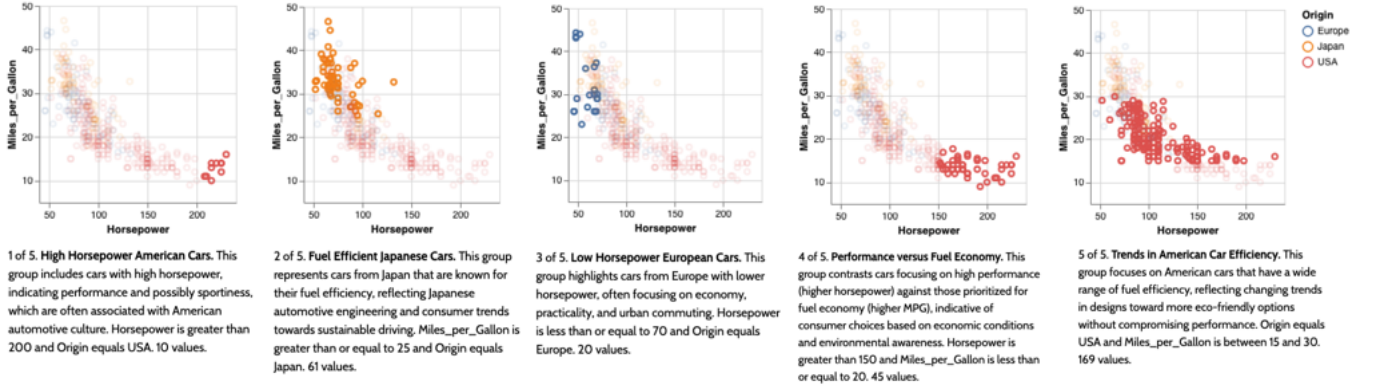


Figure 1. Data highlights generated from an example cars dataset using the *semantic scaffolding* technique. Each highlight is shown with a visualization of its query predicate, demonstrating that highlights correspond to areas of interest in the data.

mation at the expense of descriptive statistics or the data values themselves [13, 5]. These two factors require screen reader interfaces to strike a balance between making groupings available without overly prescribing a reader’s interpretation. In a study with 15 blind and low-vision users, we found that textual scaffolding was able to accelerate BLV readers’ understanding of a dataset’s real-world meaning, but that readers were often critically aware of AI’s influence on their interpretation process.

2 Related Work

2.1 Accessible Textual Data Representations

To make data visualizations accessible to screen reader users, a designer must provide descriptions that can be read as text-to-speech. Because conventional static alt text does not afford data exploration at varying levels of detail comparable to strategies sighted readers employ, researchers have turned to *structured textual descriptions* [27], which enable screen reader users to navigate along a hierarchy and move between overview and detail with textual descriptions. Accessible visualization systems that incorporate structured textual descriptions include Olli [3], Data Navigator [6], VizAbility [7], Chart Reader [25], and Umwelt [28]. This work introduces a technique for augmenting textual structures with LLM-generated groupings, and demonstrates the technique via extensions to Olli.

2.2 Data Narratives and Interpretation

Visualization researchers have used the term “narrative visualization” to describe the use of visualizations to tell stories with data [20]. Researchers studying narrative visualization have noted that the interpretation of data is not self-evident; rather, it is shaped by design and communication strategies that incorporate additional context and guide a reader through an interpretation process. For instance, “enhancing structure and navigation” and “providing controlled exploration” as important data storytelling techniques [23].

For users of textual data representations, textual annotation and interactivity (e.g. navigation) are particularly important factors that influence readers’ experience of the data. To understand the different kinds of information that authors typically include in textual descriptions or annotations, Lundgard and Satyanarayan [13] define a four-level model of semantic content in descriptions: (L1) elemental and encoded details like chart types and labels; (L2) statistical

and relational data such as outliers and correlations; (L3) perceptual and cognitive insights into trends and exceptions; and (L4) contextual and domain-specific knowledge. When they ran a study to find out what content is most useful, they found that BLV users’ preferences diverged from sighted users in a significant way. Where sighted readers rated L4 content the most useful, BLV users found it less useful because they wanted “to have the time and space to interpret the numbers for myself before I read the analysis” [13]. So while domain-specific information can be useful in descriptions, they also influence a reader’s takeaway and their feeling of agency.

Our work draws on this work to introduce interface designs for guiding a reader’s navigation through a textual data representation using domain-specific information.

2.3 Generating Textual Descriptions with Language Models

Natural language generation of textual descriptions for data visualization is an area of research that has received renewed attention due to recent advances in large language models (LLMs). There are generally two interface design approaches that these systems have taken. First are systems that provide a chat-like interface with which a user can query a description by inputting a question in natural language (often known as chart question answering systems) [11, 21, 7]. Second are systems that generate standalone summary descriptions of charts, similar in format to conventional alt text written by humans. This work includes systems like Chart-to-Text [15] and DataTales [24]. The advantage of using LLMs to generate descriptions is that they can automatically incorporate contextual and domain-specific information — also known as L4 semantic content [13] — into descriptions [12]. However, a critical limitation of this approach is the possibility that generated descriptions can contain errors, including hallucinations [10]. But even if generated captions were correct all of the time, there would still be drawbacks to existing approaches. For instance, Choe et al. noted that users sometimes become over-reliant on the LLM in a chart question answering system instead of developing their own interpretations of the data [5]. Similarly, summarization-based approaches have the same limitations that conventional alt texts do; namely, that they lack affordances for information granularity and limit users’ ability to conduct self-guided exploration [27]. In our work, we introduce a new technique — distinct from Q&A or summarization — for using LLMs to incorporate

domain-specific information. We use an LLM to generate semantically meaningful groupings, and use that output to support structural navigation [27].

3 Semantic Scaffolding: Identifying, Explaining, and Formalizing Meaningful Data Groupings

Semantic scaffolding is a technique for using domain-specific information from a large language model (LLM) to guide a reader’s understanding of a dataset’s meaning. We engaged in an iterative co-design process involving a blind co-author in which we prototyped methods for incorporating domain-specific information into a user interface for data exploration. This prototyping process revealed different types of user needs that semantic scaffolding could address, which we explore via two types of interface elements: *semantic bins* and *data highlights*.

3.1 Semantic Bins

Binning is a common operation for analyzing and communicating data that involves dividing a field into equally-sized intervals that cover the extent of the field’s data values. Most data visualizations implicitly use binning to generate axis ticks, and accessible textual data representations frequently use binning to structure a screen reader’s navigation through data.

The conventional approach to binning does not take into account domain-specific information; computing equally-sized bins is a function that can be applied to any dataset. However, lack of semantic information could make it more difficult for a reader to understand and contextualize a field’s data values. For example, in a dataset about cars, a bin function might segment the `miles_per_gallon` field by equally-sized increments of 10. However, a reader might not know what range of values is considered a low vs. high fuel efficiency, or whether an increment of 10mpg represents a large or small difference in fuel efficiency. As a result, it might be difficult for them to map the numbers onto their subjective understanding of fuel efficiency.

Semantic bins are groupings that use domain-specific information to segment a field into higher-level intervals and categories that express a dataset’s meaning. Figure 2 shows an example usage of semantic binning applied to fields with a variety of Vega-Lite measure types [1] (temporal, quantitative, and nominal). For a continuous (e.g., temporal, quantitative) field, we prompt an LLM to re-bin the field, specifying in the prompt that bins should be non-overlapping intervals that cover the extent of the data. In the figure, the semantic bins map onto how a reader might make sense of years (via historical periods corresponding to agricultural practices), and amount of wheat yield (via levels from low to high relative to the typical or expected yield) (Figure 2A). For a categorical field, we prompt an LLM to group categories into higher-level groupings, specifying in the prompt that the groupings should be mutually exclusive and exhaustively cover all categories. The figure example takes the names of barley varieties and groups them into meaningful high-level categories, such as “Heritage Varieties” or “Modern Breeds” (Figure 2B).

3.2 Data Highlights

Data highlights are groupings of data records along criteria that correspond to a real-world interpretation. In contrast to a semantic bin, which is defined by a predicate involving only one field, a data highlight’s predicate can involve multiple fields to select a subset of data records. For example, each data highlight in Figure 1

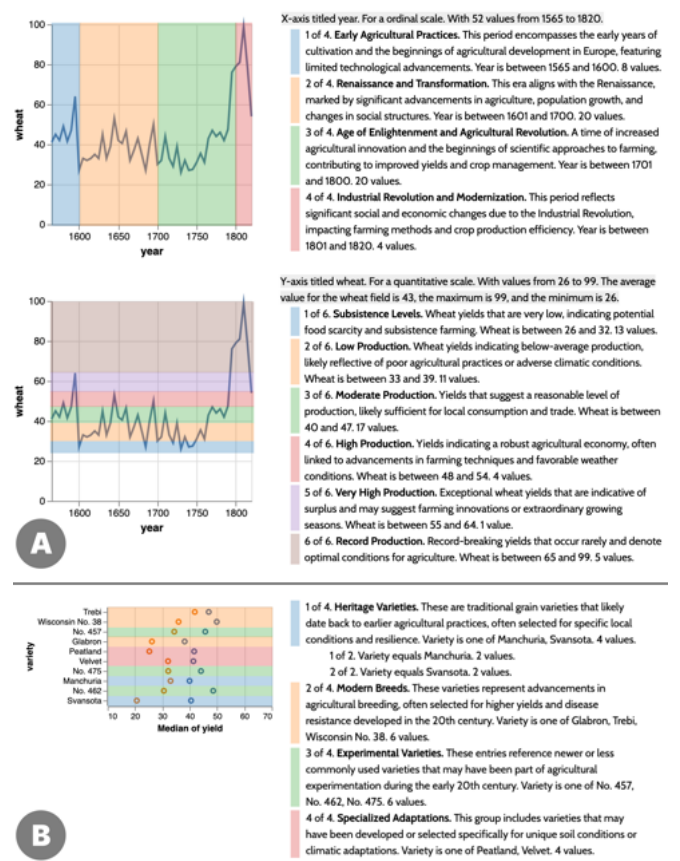


Figure 2. Semantic binning using an example wheat and barley datasets. A) the `year` field (temporal) is binned into historical periods, and `wheat` yield (quantitative) is binned into levels of low to high production. B) `barley` variety (nominal) is grouped into higher-level variety types.

represents a semantically-meaningful subset of cars in the example car dataset, with a predicate that involves two or more of the fields `Horsepower`, `Miles_per_Gallon`, and `Origin`.

Data highlights are akin to visual annotations for readers, which are a technique that designers frequently use to emphasize and draw attention certain parts of the data [18]. Indeed, data highlights have the same components as many instances of visual annotation: a defined subset of data records, and a semantically-meaningful explanation. However, we think of data highlights as independent of any specific visual representation. In our examples, we convey data highlights both as a conditional encoding in a visualization (e.g. visually annotating the included data points using color or opacity), and as descriptions in a textual structure (Figure 1).

Data highlights are designed to help lay readers understand a dataset even if they lack prior knowledge about the data domain. For example, in the cars dataset from Figure 1 a reader might observe the inverse relationship between `Horsepower` and `Miles_per_Gallon` but lack the context to know why there might be a tradeoff between the two, or how cultural factors between the three `Origin` values contribute to differences. In Figure 1, each data highlight focuses on a different region of the chart, and connects the selected data points with a real-world explanation. For example, the third data highlight in the figure is called “Low Horsepower European Cars,” and explains the tendency for European cars to have lower horsepower in terms of economy, practicality, and urban com-

muting. The highlight includes a query predicate defining the grouping’s inclusion criteria in terms of the `Horsepower` and `Origin` fields, allowing the reader to connect the explanation to concrete values in the dataset.

3.3 Limitations

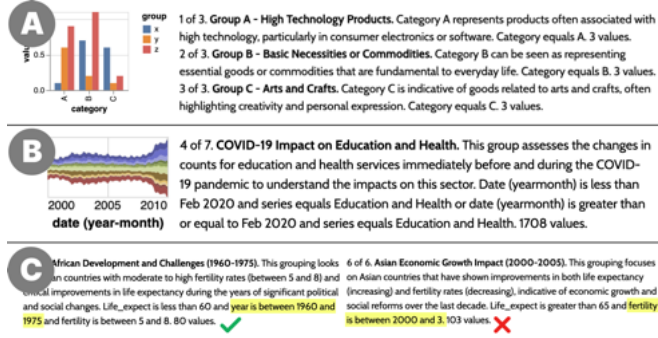


Figure 3. Examples of LLM errors encountered in our design process. A) LLM hallucinates dataset semantics when there is not enough information. B) LLM includes information that is outside of the dataset’s context. C) LLM generates incorrect query predicate.

Because semantic scaffolding is a technique that relies on LLMs, it is limited by the kinds of errors LLMs tend to make. Here, we provide examples of errors we observed in our prototyping process.

Hallucinated dataset semantics. When a dataset does not have meaningful field labels, the LLM might hallucinate a domain for the data. In [Figure 3A](#), the data has the generic field labels of `category` and `value`. Even though there’s no information about what these fields mean, the LLM interprets them as consumer products.

Social context inappropriate to dataset. The LLM may incorporate domain-specific information that is not necessarily factually wrong, but is outside of the scope of the dataset. [Figure 3B](#) shows a data highlight referencing the COVID-19 pandemic, which started in 2020, for an unemployment dataset that only covers 2000–2010.

Incorrect query predicates. [Figure 3C](#) demonstrates an error where the LLM identified a reasonable grouping, but was unable to formulate a correct query predicate that matches the explanation. The figure shows one correct predicate and one incorrect predicate generated from the same data. While the `fertility` between 2000 and 3 term is syntactically valid, it is semantically incorrect.

Discussion of limitations. In our prototyping process, we explored recent techniques for LLM output validation, such as SymGen [\[8\]](#). However, we found that most data-to-text validation techniques were not appropriate for our use case, because we are not merely asking the LLM to summarize a dataset. In our case, our goal is to leverage the LLM’s knowledge that is external to the dataset. As a result, validating this information requires either access to an external knowledge base, or use of a proxy like the model’s internal confidence [\[26\]](#). Therefore, because this is still an active area of research, we were not able to incorporate reliable technical approaches to validating the output of semantic scaffolding at this time. However, in our user testing, we surfaced insights on user behaviors and strategies for validating and thinking critically about LLM output ([section 5](#)).

4 Implementation

We implemented semantic scaffolding by prompting gpt-4o-mini with instructions to create and return data groupings, attaching a full dataset to the prompt. To ensure that we receive a structured response from the LLM, we defined a data structure that specifies our expected response format. Each grouping identified by the LLM is returned as a data structure with three properties: a name that summarizes the content of the grouping; a longer explanation of the grouping’s meaning; and a query predicate that defines the criteria for a data point to be included in the grouping. For compatibility with existing visualization tools, the query predicate is specified in Vega-Lite’s existing predicate syntax [\[19\]](#). [Figure 4](#) contains a definition of the data structure in TypeScript syntax.

```
type LLMResponse = {
  groups: SemanticScaffold[];
};

type SemanticScaffold = {
  name: string;
  explanation: string;
  predicate:
    ↳ LogicalComposition<FieldPredicate>;
};

{
  "name": "AAPL Price Surge During the Tech Boom",
  "explanation": "This group focuses on Apple's stock price during the late 2000s and early 2010s, highlighting its rise in value matched with the smartphone revolution and innovation of products like the iPhone.",
  "predicate": {
    "and": [
      { "field": "symbol", "equal": "AAPL" },
      { "field": "price", "gte": 150 },
      { "field": "date", "range": [ "2008-08-31", "2012-12-31" ] }
    ]
  }
}
```

Figure 4. Type definitions and example for semantic scaffolding

5 Evaluation: User Study

To evaluate semantic scaffolding, we conducted a user study with 15 blind and low-vision (BLV) participants who use screen readers. The goal of this evaluation was to understand how our technique affects readers’ data exploration and interpretation. In this section, we describe the methods we used to design our evaluation, and report survey and interview results.

5.1 Methods

The study involved 100-minute Zoom interviews with 15 BLV participants exploring three prototype implementations of semantic scaffolding in Olli [\[3\]](#). We began each interview with a brief 10 minute

introduction to learn more about the user’s interactions with data, data visualizations, and large language models. To familiarize participants with navigation in Olli, we did a 15 minute introduction of the existing baseline version of Olli. Then, we had users explore the *data-first* and *highlights-first* prototypes (described below), each for 20 minutes. For each prototype, we guided participants through the prototypes and asked participants to explore the data and comment on their thought process. Then, we had users complete the prototype’s corresponding Likert survey. In the final 15 minutes, we discussed users’ overall takeaways and conducted a post-study survey.

Prototypes. In our study, we used the default version of Olli as a baseline, and compared it to two prototypes with semantic scaffolding. These two prototypes explored different ways to present semantic scaffolds in a screen reader interface. One prioritizes showing additional domain-specific information first, while the other prioritizes direct access to the data first. We created these two variations because we were interested in whether readers would prefer to have highlights first to orient themselves, or whether they would rather explore the data directly before zooming back out for additional context. For us, this question was connected to the idea of agency — for instance, would data highlights make a reader feel like they were being told what to think rather than drawing their own conclusions? For both prototypes, we presented a scatterplot but with a different dataset each time so we could understand how a prototype helped them make sense of new datasets. We chose scatterplots because they are typically considered difficult to read for screen reader users, and because sighted readers tend to identify groupings in scatterplots using visual features like clustering. For the user study, we pre-generated and hand-curated LLM responses to ensure that they made sense, were the same for each participant, and would load quickly.

- **Olli baseline.** We use the existing version of Olli as a baseline and to familiarize participants with textual structures. This baseline uses an example movies dataset with fields `US_Gross`, `IMDB_Rating`, and `MPAA_Rating`.
- **Data-first prototype.** This prototype prioritizes giving readers a chance to explore the data before adding additional context. It conserves the original Olli structure as much as possible, and appends highlights to the end of the encodings level of the Olli tree view. This prototype uses an example penguins dataset with fields `Flipper_Length (mm)`, `Body_mass (g)`, and `Species`.
- **Highlights-first prototype.** This prototype prioritizes giving an overview with data highlights first, before opening up further exploration in the data. Data highlights are presented as a subtree of annotations, prepended to the encodings level of the Olli tree view. This prototype uses an example cars dataset with fields `Horsepower`, `Miles_per_Gallon`, and `Cylinders`.

Participants. We recruited 15 blind and low-vision participants by sending a participant call to a blind programmers’ community mailing list and reaching out to a local BLV community group. Each participant was compensated \$60 for 100 minutes. To protect participant privacy, we’ve included anonymized and aggregated demographic information to provide background context. The majority (67%) of our participants were totally blind (n=10), while 27% identified as low-vision with some light perception (n=4) and 7% of participants did not respond (n=1). More than half (60%) of our participants have been blind since birth. 53% of participants were JAWS users (n=8) and 47% were NVDA users (n=7), relatively consistent with screen reader statistics [2]. Demographically, 80% of participants use he/him pronouns (n=12) and 20% of participants

use she/her pronouns (n=3). Participants were based across multiple continents, including North America, Europe, and Asia. Participants self-reported their ethnicities (Asian, Black/African, Caucasian/white, Hispanic/Latinx, Other), represented a diverse range of ages (20–50+), and had a variety of educational backgrounds (high school through to undergraduate and graduate school). With the exception of one participants that did not respond, almost all participants (n=14) self-reported as slightly, somewhat, or moderately familiar with statistical concepts. 13 participants self-reported as slightly, somewhat, or moderately familiar with data visualization methods and 2 participants did not respond. Participants reported a high variety of frequency interacting with data or visualizations, from 1-2 times/year to 3 or more times/week. Most participants (n=9) reported using data analysis tools or visualizations either outside of their professional work or sometimes, but 2 participants reported data analysis tools or visualizations being an important part of their workflow. 4 participants reported rarely use data analysis tools.

5.2 Quantitative Results

We designed a Likert survey to understand participants’ preferences across the highlight-first and data-first prototypes. Participants responded on a five point scale where 1 = Very Difficult to Understand / No Influence / No Additional Context / Not Effective / Never Felt the Need and 5 = Very Easy to Understand / Significant Influence / Extensive Additional Context / Extremely Effective / Always Felt the Need (Table 1).

The highlights-first prototype consistently scored higher in terms of understandability for both highlights and bins, indicating that users found this prototype more intuitive and easier to use. Despite this, the influence of highlights and bins on data interpretation was similar between the two prototypes, suggesting that while the highlights-first prototype was easier to understand, it did not significantly change how users interpreted the data compared to the data-first prototype. However, participants felt the highlights-first prototype was more effective in explaining and justifying its data highlights and bins, as reflected by higher scores in that category. Users also felt a slightly greater need to double-check the accuracy of data highlights and bins in the highlights-first prototype, though this difference was minimal. These results suggest that participants generally preferred the highlights-first prototype, which offers better usability and clarity. However, these findings are best contextualized with our qualitative findings in the next section, especially around participants’ trust in AI and desire for agency in interpretation.

5.3 Qualitative Results

Several themes emerged during our observation of participant interaction with the prototypes, and in the interviews that followed. To analyze our qualitative data, authors Zong, Pedraza Pineros, and Hajas independently conducted qualitative coding using interview transcripts, grouping participant quotes into thematic categories. Then, authors synthesized these themes into the findings reported in this section.

5.3.1 Semantic scaffolding helps readers understand and contextualize data

When encountering a new dataset, readers need to understand the data’s real-world meaning first. P1 illustrated this point nicely, saying “the numbers don’t mean much until I know a little bit about what

Table 1. Rating scores for each prototype (Data-first Prototype, Highlights-first Prototype) on a five point Likert scale. Median scores are shown in boldface, averages in brackets, standard deviations in parentheses.

Prompt: After understanding how the [prototype] works...	Data-first	Highlights-first
How understandable were the data highlights in the prototype?	4 [4.13] (0.64)	5 [4.53] (0.64)
How understandable were the bins in the prototype?	4 [4.00] (1.20)	5 [4.67] (0.49)
How much influence did the data highlights have on your interpretation of the data?	4 [3.60] (1.30)	4 [3.33] (1.45)
How much influence did the bins have on your interpretation of the data?	3 [3.13] (1.25)	3 [3.00] (1.36)
How much additional context did the data highlights provide about the data?	4 [3.93] (1.10)	4 [3.60] (1.24)
How much additional context did the bins provide about the data?	4 [3.47] (0.92)	4 [3.47] (0.83)
How effectively did the prototype explain and justify its data highlights and bins?	3 [2.93] (0.96)	4 [3.73] (0.88)
How often did you feel the need to double-check the accuracy of the data highlights and bins?	2 [2.33] (1.18)	2 [2.47] (1.06)

I’m reading about.” However, accurately understanding the dataset’s semantics can be difficult for readers. We found that participants were able to quickly identify what the data refers to using semantic scaffolds even if their initial understanding was incorrect. For instance, participants had many different initial guesses about the prototypes’ data semantics. Based on the `Flipper Length (mm)` and `Body Mass (g)` fields, some participants initially thought the penguins dataset referred to fish (P1), seals (P2), or dolphins (P3). However, all these participants correctly updated their understanding to reflect information about penguins found in the data highlights. Accelerating this process of identifying data semantics can be very important to a reader’s experience: as one participant put it, “I don’t feel dumb looking at this data” (P4).

One participant contrasted semantic scaffolding positively with prior AI-based tools they had used. P9, who had used chart question-answering tools in the past, noted that “sometimes you don’t know what questions to ask” when you are not familiar with the data. Further, they said “sometimes its unfair to expect you to ask the right questions because you don’t have a clue of [...] what the data is about.” Semantic scaffolding can address this “cold start” problem when participants start with limited prior knowledge.

5.3.2 Reshaping textual structures using semantic scaffolding helps with navigation and wayfinding

Participants found semantically meaningful groupings helpful for navigating through the data. For example, P5 said that “the groupings keep it in an organized fashion so that I get a more orderly presentation” of the data. Others similarly noted that groupings help them not “wander uselessly around,” (P6) and not “have to look for things random” (P4) and instead explore within a relevant space. These comments echo previous findings that screen reader users benefit from interfaces that provide a *bounded space* [27] that helps a user locate themselves within a smaller space with known boundaries.

Other participants also found semantic scaffolds helpful for identifying where they could navigate to find more relevant information. For instance, P8 said that semantic bins helped “frame my expectations in advance to have some information of what the legend is.” For them, the description of the semantic bin was providing *information scent* [17] about what could be found when navigating further into that grouping. This finding reinforces the value of nodes that compare across multiple data entries or fields, representing meaningful

subsets of data, and do not require manual specification of filters by a user.

5.3.3 Readers critically appraise semantic scaffolds using their prior knowledge

Semantic scaffolds prompted participants to think about data in semantically meaningful ways, incorporating their own prior knowledge. For example, engaging with LLM-generated categorizations of data prompted participants to imagine other ways of thinking about the data’s meaning, and other possible categorizations. When reading the categorization of cars in the example dataset, P2 noted that “one of the things that’s missing is work vehicles, or pickup trucks.” In other words, there was a category that they expected to see that was not present. Similarly, P8 noted that “car manufacturers don’t [categorize cars with] engine sizes, they just base it on fuel economy,” and suggested “utility” as a potential additional category. In this case, the participant had additional context or expertise that made the LLM’s categorization insufficient for them, even if it was not technically incorrect.

Sometimes, participants did not necessarily disagree with the content of semantic groups, but were skeptical of arbitrariness and asked for justification. For example, P1 questioned the cutoffs for semantic bins, saying, “I can hear that it’s grouping cars by cylinders and horsepower, but why are these the breaks? Why 101 to 200, or why eight cylinders? It seems arbitrary.” Similarly, P11 questioned the subjectivity of certain categories: “I don’t know how arbitrary the cutoff is. Like, what would you say a gas guzzler is? That’s maybe arbitrary, maybe not. I don’t know.” Even when participants lacked domain knowledge to fully evaluate explanations, they were still aware of potential limitations in the groupings.

A person’s cultural context can also be a factor, as we observed when P6 pointed out that “the values don’t correspond to my knowledge because I use different units.” Because the data used miles and gallons, which are units that not every country uses, the explanation was less useful for them.

However, people appreciated when the descriptions matched their existing knowledge, whatever level of knowledge that may be. Many applied the standard of “it sounds plausible” when evaluating groupings (P3, P9). For participants who did have more prior knowledge, they felt like they were still able to learn something new. For instance, P8 said of a semantic binning that they “know a little about

cars,” but “without it telling me I would have no idea that this range is economy / mid range.” They felt that they were provided with useful information they still would not have known otherwise.

5.3.4 *Trust in AI-generated descriptions is situational*

Participants had a range of attitudes about their trust in AI-generated information. Some were fairly confident in the information they were reading in the semantic scaffolds, as long as it seemed plausible. For example, participants remarked, “I didn’t hear anything outlandish” (P3), and “I assume that the domain experts would agree with those descriptions” (P11). Even participants that were fairly aware of LLM hallucinations said that, even if they might double-check later, they would generally take the descriptions at face value (P1, P3, P5). For example, P3 said, “I might go to Google and re-read on these things but at least I will take the first impressions here.”

On the other hand, some participants expressed more concerns about the risks of incorrect descriptions. Others said that “hallucination is very dangerous” (P6), noting that LLMs “sometimes give you really really false data with high confidence” (P13). These participants felt they were more likely to verify information with external sources. A consistent factor in people’s willingness to double-check output is how important or high-stakes their data analysis felt. Many said that they would be more likely to check against other sources if they were looking at data for professional purposes (P1), making a presentation (P2), or if it was otherwise really important (P6, P8).

Participants employed a variety of strategies to manage trust and verification of LLM-generated information. They mentioned using Google or their own knowledge (P3), finding another non-automated source (P5), or refreshing the LLM output to see if it remained consistent (P1). In our study, for example, participants used the data table to compare against the description of the grouping’s query predicate, to check if it “matches the description” (P11). Overall, many participants shared a pragmatic attitude. For instance, P6 said, “you have to take what is good from it, but with a grain of salt.”

5.3.5 *AI-generated descriptions might risk reducing readers’ agency to interpret the data*

Participants’ preferences between the highlights-first and data-first prototypes often boiled down to how much they wanted to rely on semantic scaffolding to understand the data. Participants generally preferred the highlights-first prototype, echoing previous findings about the importance of an “overview first, then detail” strategy in data exploration [22, 27]. For example, P3 preferred the highlights first because “it gives an overall picture / what to expect before I dig into the data.” However, P14 dissented, saying that “I’m skeptical of people interpreting my data for me, [so] I like the facts first.”

These differing preferences were emblematic of different perspectives on how the AI-generated descriptions were influencing reader interpretation. The more common view was that semantic scaffolding has a negative influence on reader agency (P2, P6, P10, P14). For instance, P2 said, “I feel like my conclusions are being shaped. If someone first gives you the conclusion, it kind of shapes your perception. If you are first given the raw data, you try to make sense for yourself and after that you can compare it to the conclusions provided.” Similarly, P14 drew a distinction between interpretative and factual content in descriptions, saying, “I want the AI to separate interpretation from facts. ‘Body mass’ is fact, ‘healthy body mass’ is getting into interpretation.” In this instance, the participant felt a semantic bin titled “healthy body mass” and associated with a specific

range was imposing a subjective interpretation on the data, rather than letting the reader decide what range is considered healthy. These concerns are consistent with prior findings that unlike sighted readers, BLV readers find interpretive content less useful in descriptions of data [13]. P6 felt that they would get “spoiled” by the tool, saying “I would lose the tendency of exploring and rely on the highlights too much.” This concern is consistent with prior findings too, that users sometimes become over-reliant on LLM agents in data analysis, leading to less overall engagement with data [5].

However, there were also a number of participants who felt that semantic scaffolding had a positive influence on their interpretation process (P5, P11). For instance, P11 said of data highlights, “It’s giving me ideas about what questions to ask or what relationships are worth looking at.” Similarly, P5 said that semantic scaffolds “[give] me a sense of confidence now that I have access to information that I didn’t have before. This information is presented in a means that I can obtain it, digest it and work with it in my own way.” In this case, the participants felt that data highlights were getting them started on asking their own followup questions about the data — rather than foreclosing further interpretation.

5.3.6 *Readers expected semantic scaffolds to augment the original data*

Multiple participants expected semantic groupings to be reflected in the dataset itself, and were confused when that did not happen. For example, P10 said, “You gave me this description here of, you know, sports cars, muscle cars, sedans.” When they then drilled down into the data table view, they were “a little bit frustrated that you don’t specify which one is a sedan and which one is a muscle car.” They then elaborated that they expected the AI to “add that information as an extra column to the table” (P10). This suggests that participants did not necessarily distinguish semantic scaffolds from the data itself, because they both describe the same real-world domain. Future work could potentially explore semantic scaffolding in conjunction with data augmentation [4] to synchronize these two views, or otherwise manage this mismatch in expectations.

5.3.7 *BLV readers also want sighted readers to have semantic scaffolding, to preserve common ground*

For many participants, the use of semantic scaffolding made them reflect on situations where they are collaborating with sighted people. For example, participants noted that semantic scaffolding provides “maybe more than a sighted person would get from such a graphic because it gets you info about the relevance of the values with knowledge you don’t have” (P6) and that this “feels weird when working with sighted people” (P8). Participants generally felt that it was important to maintain parity of information access with sighted people. For example, P10 asked, “does the visual person get the same info?” Prior work has emphasized the importance of common ground between mixed-ability collaborators [28], and participants’ concerns reflected the idea that it was important to them to participate in and lead conversations about data even in spaces that have a majority of sighted people. P14 explained, “whenever you can put sighted and blind individuals on the same playing field with equal footing, you’ve accomplished something that’s really marvelous. [...] For us to be able to be literally on the same page, it’s helpful because we’re looking at the very same information.” For P14, accessibility is a two-way street: “[blind people] have to accommodate [visual people] in a way.” It was important to them to include sighted collaborators

in their data analysis process, even when using assistive tools like semantic scaffolding. These discussions led us to reflect on the distinction in our work between *disability inclusion for BLV people* and *making data understandable for lay readers*, two ideas of accessibility that can be complementary and mutually-reinforcing.

6 Conclusion

In this paper, we introduced semantic scaffolding: a technique for using an LLM to generate meaningful groupings of data to aid lay readers in understanding unfamiliar datasets. With semantic scaffolding, readers benefit from domain-specific knowledge that is incorporated into descriptions and navigational structures. We instantiated this approach in Olli, and observed that users find semantic scaffolds valuable for forming an initial understanding of a dataset while applying their prior knowledge to evaluating the trustworthiness of descriptions. Our work suggests a use for LLMs to augment user interfaces with domain-specific affordances for interpretation.

ACKNOWLEDGEMENTS

This work was supported by NSF #2341748, MIT-SERC, and MIT-Google.

REFERENCES

- [1] Type | Vega-Lite, 2023.
- [2] Screen Reader User Survey #10 Results, 2024.
- [3] Matthew Blanco, Jonathan Zong, and Arvind Satyanarayan, ‘Olli: An Extensible Visualization Library for Screen Reader Accessibility’, in *IEEE VIS Posters*, (2022).
- [4] Dylan Cashman, Shenyu Xu, Subhjit Das, Florian Heimerl, Cong Liu, Shah Rukh Humayoun, Michael Gleicher, Alex Ender, and Remco Chang, ‘CAVA: A Visual Analytics System for Exploratory Columnar Data Augmentation Using Knowledge Graphs’, *IEEE Transactions on Visualization and Computer Graphics*, **27**(2), 1731–1741, (February 2021).
- [5] Kiroong Choe, Chaerin Lee, Soohyun Lee, Jiwon Song, Aeri Cho, Nam Wook Kim, and Jinwook Seo, ‘Enhancing Data Literacy On-demand: LLMs as Guides for Novices in Chart Interpretation’, *IEEE Transactions on Visualization and Computer Graphics*, 1–17, (2024).
- [6] Frank Elavsky, Lucas Nadolskis, and Dominik Moritz, ‘Data Navigator: An Accessibility-Centered Data Navigation Toolkit’, *IEEE Transactions on Visualization and Computer Graphics*, 1–11, (2023).
- [7] Joshua Gorniak, Jacob Ottiger, Donglai Wei, and Nam Wook Kim, ‘VizAbility: Multimodal Accessible Data Visualization with Keyboard Navigation and Conversational Interaction’, in *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST ’23 Adjunct, pp. 1–3, New York, NY, USA, (October 2023). Association for Computing Machinery.
- [8] Lucas Torroba Hennigen, Zejiang Shen, Aniruddha Nrusimha, Bernhard Gapp, David Sontag, and Yoon Kim, ‘Towards Verifiable Text Generation with Symbolic References’, in *First Conference on Language Modeling*, (2024).
- [9] Jessica Hullman and Nick Diakopoulos, ‘Visualization Rhetoric: Framing Effects in Narrative Visualization’, *IEEE Transactions on Visualization and Computer Graphics*, **17**(12), 2231–2240, (December 2011).
- [10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung, ‘Survey of Hallucination in Natural Language Generation’, *ACM Computing Surveys*, **55**(12), 1–38, (December 2023).
- [11] Ecem Kavaz, Anna Puig, and Inmaculada Rodríguez, ‘Chatbot-Based Natural Language Interfaces for Data Visualisation: A Scoping Review’, *Applied Sciences*, **13**(12), 7025, (January 2023). Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.
- [12] Jiho Kim, Arjun Srinivasan, Nam Wook Kim, and Yea-Seul Kim, ‘Exploring Chart Question Answering for Blind and Low Vision Users’, in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, pp. 1–15, New York, NY, USA, (April 2023). Association for Computing Machinery.
- [13] Alan Lundgard and Arvind Satyanarayan, ‘Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content’, *IEEE Transactions on Visualization and Computer Graphics*, **28**(1), 1073–1083, (January 2022).
- [14] Tamara Munzner, *Visualization Analysis and Design*, A K Peters/CRC Press, New York, October 2014.
- [15] Jason Obeid and Enamul Hoque, ‘Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model’, in *Proceedings of the 13th International Conference on Natural Language Generation*, eds., Brian Davis, Yvette Graham, John Kelleher, and Yaji Sripada, pp. 138–147, Dublin, Ireland, (December 2020). Association for Computational Linguistics.
- [16] Evan M. Peck, Sofia E. Ayuso, and Omar El-Etr, ‘Data is Personal: Attitudes and Perceptions of Data Visualization in Rural Pennsylvania’, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, Glasgow Scotland Uk, (May 2019). ACM.
- [17] Peter Pirolli and Stuart Card, ‘Information foraging’, *Psychological Review*, **106**(4), 643–675, (1999). Place: US Publisher: American Psychological Association.
- [18] Donghao Ren, Matthew Brehmer, Bongshin Lee, Tobias Hollerer, and Eun Kyoung Choe, ‘ChartAccent: Annotation for data-driven storytelling’, in *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 230–239, Seoul, South Korea, (April 2017). IEEE.
- [19] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer, ‘Vega-Lite: A Grammar of Interactive Graphics’, *IEEE Transactions on Visualization and Computer Graphics*, **23**(1), 341–350, (January 2017).
- [20] Edward Segel and Jeffrey Heer, ‘Narrative Visualization: Telling Stories with Data’, *IEEE Transactions on Visualization and Computer Graphics*, **16**(6), 1139–1148, (November 2010).
- [21] Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang, ‘Towards Natural Language Interfaces for Data Visualization: A Survey’, *IEEE Transactions on Visualization and Computer Graphics*, **29**(6), 3121–3144, (June 2023).
- [22] Ben Shneiderman, ‘The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations’, in *The Craft of Information Visualization*, eds., Benjamin B. Bederson and Ben Shneiderman, Interactive Technologies, 364–371, Morgan Kaufmann, San Francisco, (2003).
- [23] Charles D Stolper, Bongshin Lee, Nathalie Henry Riche, and John Stasko, ‘Emerging and Recurring Data-Driven Storytelling Techniques: Analysis of a Curated Collection of Recent Stories’, Technical Report MSR-TR-2016-14, Microsoft Research, (2016).
- [24] Nicole Sultanum and Arjun Srinivasan, ‘DataTales: Investigating the use of Large Language Models for Authoring Data-Driven Articles’, pp. 231–235. IEEE Computer Society, (October 2023).
- [25] John R Thompson, Jesse J Martinez, Alper Sarikaya, Edward Cutrell, and Bongshin Lee, ‘Chart Reader: Accessible Visualization Experiences Designed with Screen Reader Users’, in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, Hamburg Germany, (April 2023). ACM.
- [26] Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn, ‘Fine-Tuning Language Models for Factuality’, in *The Twelfth International Conference on Learning Representations*, (2024).
- [27] Jonathan Zong, Crystal Lee, Alan Lundgard, JiWoong Jang, Daniel Hajas, and Arvind Satyanarayan, ‘Rich Screen Reader Experiences for Accessible Data Visualization’, *Computer Graphics Forum*, (2022).
- [28] Jonathan Zong, Isabella Pedraza Pineros, Mengzhu (Katie) Chen, Daniel Hajas, and Arvind Satyanarayan, ‘Umwelt: Accessible Structured Editing of Multi-Modal Data Representations’, in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI ’24, pp. 1–20, New York, NY, USA, (May 2024). Association for Computing Machinery.

The Development and Evaluation of a Hybrid ViT-LSTM for Word-Level Sign Language Recognition

Charlotte Richardson¹ and Ronald R. Grau²

Abstract. This paper presents a hybrid Vision Transformer-Long Short-Term Memory (ViT-LSTM) model for word-level sign language recognition (SLR) in low-resource settings. A pretrained ViT backbone extracts per-frame spatial features, which are passed to a Bidirectional LSTM to model temporal dynamics across gesture sequences. A learned attention mechanism produces a context-aware representation for classification. The architecture is evaluated first on MMLAuslan, a large-scale Australian Sign Language (Auslan) dataset, where domain-proximate ViT pretraining achieved 71% validation accuracy, and 55% test accuracy on 35 alphanumeric signs versus 44% for a general-purpose baseline. Trained on a linguistically motivated 116-gloss vocabulary, the model reached 77% validation accuracy and 78% test accuracy using approximately 1,400 training samples. A key contribution is a cross-lingual transfer pipeline from Auslan to BSL, exploiting shared phonological structure between the two languages. BSL fine-tuning uses progressive backbone unfreezing and separate learning rates to preserve transferable representations. Experiments on BSL training using transfer learning show accuracy improvement over a BSL-only baseline. Attention map analysis confirms the model attends to task-relevant hand regions rather than background artefacts. Future work will incorporate pose estimation and hand landmarks, with the longer-term goal of real-time learner feedback for classroom use. The aim is to ensure that SLR development actively supports equitable participation for DHH learners rather than reproducing existing resource inequalities.

1 INTRODUCTION

For Deaf and hard-of-hearing (DHH) learners, communication barriers in hearing-centred educational environments remain a persistent challenge. Automatic sign language recognition (SLR) capable of operating in real-time could address this directly, supporting classroom communication assistance, independent vocabulary acquisition, and automated feedback on signing accuracy [5]. Access to such tools is not merely beneficial but a prerequisite for meaningful participation in inclusive education [2]. Yet SLR research has been highly uneven across languages. The vast majority of work targets well-resourced languages such as American Sign Language (ASL), while British Sign Language (BSL) remains largely unsupported by modern deep learning approaches [11, 14]. This imbalance reflects a structural inequality in AI-supported education in which the availability of assistive technology is constrained by the scarcity of large-scale annotated datasets for minority sign languages.

This paper presents a hybrid Vision Transformer–Long Short-Term Memory (ViT-LSTM) architecture for word-level, isolated SLR, designed explicitly for low-resource settings. Sign languages are spatiotemporal systems [12] in which individual signs are spatially structured within each frame while meaning is also conveyed through movement, timing, and the dynamic evolution of handshape across a temporal sequence [9]. Purely convolutional approaches capture spatial detail but struggle with long-range temporal dependencies [4], while recurrent models handle sequences well but lack strong spatial inductive biases [6]. The proposed hybrid model addresses both challenges within a single architecture.

The system was evaluated in two stages on the MM-WLAuslan dataset [16], a large and balanced Australian Sign Language (Auslan) corpus, before being adapted to a custom BSL dataset via cross-lingual transfer learning. The central contribution is a cross-lingual transfer pipeline from Auslan to BSL, exploiting the phonological proximity of the two languages [1] to compensate for the scarcity of BSL training data.

Section 2 reviews relevant background. Section 3 describes the model architecture and training strategy. Section 4 details the datasets and experimental design. Section 5 reports results. Section 6 discusses the findings, and Section 7 concludes.

2 BACKGROUND

2.1 Sign Language Recognition

SLR approaches divide broadly into sensor-based and vision-based methods. Sensor-based systems such as datagloves can achieve very low error rates but remain impractical due to cost and scalability constraints [14]. Vision-based deep learning approaches have therefore become dominant. Convolutional Neural Networks (CNNs) have achieved strong results on static sign images [14], but sign language is inherently dynamic. Most signs involve coordinated movement of both hands over time, making purely spatial models insufficient. 3D-CNNs extend to video but lack the global self-attention mechanisms needed to capture long-range spatial dependencies [4].

2.2 Vision Transformers

Introduced by Dosovitskiy et al. [4], Vision Transformers (ViTs) apply transformer self-attention directly to sequences of image patches. Unlike CNNs, ViTs integrate information across the entire image simultaneously, making them well suited to the fine-grained spatial structure of sign language, where hand position relative to the face and body carries semantic meaning. A key constraint is that ViTs typically require very large datasets to train from scratch; however,

¹ University of Sussex, email: cr464@sussex.ac.uk

² University of Sussex, email: grau@sussex.ac.uk

pretrained ViTs transfer effectively to smaller target domains, making them particularly appropriate for low-resource SLR.

2.3 Long Short-Term Memory Networks

LSTMs [6] are recurrent networks designed to model long-range temporal dependencies via gating mechanisms. Their application to SLR is well established. Sundar and Bagyammal [18] combined MediaPipe pose extraction with an LSTM to achieve 99% accuracy on ASL fingerspelling, while Huang and Chouvatut [8] coupled a ResNet spatial encoder with an LSTM to obtain 86.25% on Argentine Sign Language. These results confirm that temporal modelling is a critical component of high-performing SLR systems.

2.4 Hybrid Architectures and Transfer Learning

Chen et al. [3] demonstrated that combining a ViT feature extractor with a bidirectional LSTM for micro-expression recognition achieved 86.7% accuracy and an F1 of 86.4%, confirming the complementary nature of the two components for spatiotemporal tasks. For low-resource sign languages, Mocialov et al. [15] showed that transfer learning from a related sign language corpus produces better initial performance, faster convergence, and improved final accuracy compared to training from scratch, establishing cross-lingual transfer as a viable strategy for bootstrapping BSL recognition.

3 MODEL ARCHITECTURE AND TRAINING STRATEGY

3.1 Architecture Overview

The hybrid model comprises three components: a ViT backbone for per-frame spatial feature extraction, a bidirectional LSTM with learned temporal attention for sequence modelling, and a classification head mapping temporal representations to gloss classes. The architecture is illustrated in Figure 1.

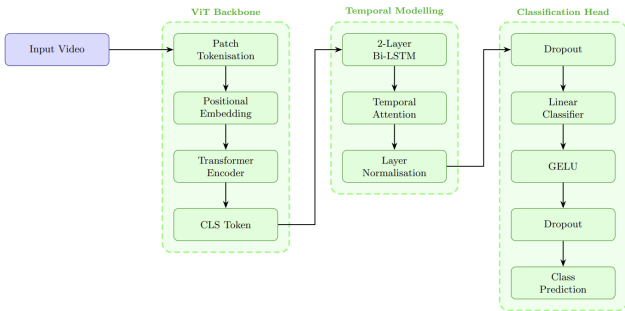


Figure 1: Hybrid ViT-LSTM model architecture.

ViT Backbone. Each input frame is processed independently by the ViT backbone. The frame is divided into 16×16 -pixel patches, linearly projected into token embeddings, and enriched with positional encodings before being passed through the transformer encoder [4]. The resulting CLS token, a 768-dimensional global representation, is used as the per-frame embedding. This compact representation avoids the complexity of pooling the full patch token sequence and provides a stable input to the temporal component. The backbone selected for all experiments was Dima806, a Google ViT

pretrained on ImageNet-21k and fine-tuned on HaGRID, a large-scale hand gesture dataset, which outperformed both a general ImageNet backbone and a human action recognition backbone in preliminary experiments.

Bidirectional LSTM and Temporal Attention. Sign gestures are dynamic and therefore individual frames are insufficient to discriminate between signs. The 20 per-frame ViT embeddings are passed to a two-layer Bidirectional LSTM (Bi-LSTM), enabling the model to integrate contextual information from both past and future frames within the sequence [3]. A learned scalar attention mechanism is applied over the Bi-LSTM output, assigning importance weights to each frame before producing a context-aware pooled representation. This is preferable to mean pooling because transitional and preparatory frames carry less discriminative information than peak sign frames [10]. The attended representation is layer-normalised before being passed to the classification head.

Classification Head. The head consists of two fully connected layers with intermediate dimensionality of 256, a GELU activation between them, and dropout applied before each layer (rates of 0.5 and 0.3 respectively) to regularise training under limited data.

3.2 Training Strategy

Training followed a two-phase progressive unfreezing strategy [7, 21]. In Phase 1, the ViT backbone was frozen and only the Bi-LSTM and classification head were updated, protecting pretrained spatial representations from large early gradient updates while allowing the temporal components to stabilise. In Phase 2, the final three transformer layers were selectively unfrozen and updated with a backbone learning rate an order of magnitude lower than the head, with a linear warmup scheduler applied over 10 epochs. Cross-entropy loss with label smoothing of 0.1 was used to prevent overconfidence [19], and AdamW was selected as the optimiser for its improved generalisation in fine-tuning settings [13].

BSL Transfer Learning Configuration. For BSL fine-tuning, the best Auslan checkpoint was loaded as the starting weights, with the classification layer reinitialised for the BSL target vocabulary. Phase 2 unfroze eight rather than three backbone layers to improve cross-lingual domain adaptation. Augmentation strength was reduced relative to Auslan training, as the Auslan backbone already encodes substantial visual invariance [17].

Input videos were represented as fixed-length sequences of 20 frames, sampled uniformly with a configurable offset to skip pre-sign pauses, and each frame was resized to 224×224 pixels.

4 DATASETS AND EXPERIMENTAL DESIGN

4.1 Datasets

MM-WLAuslan. The primary dataset was MM-WLAuslan [16], comprising 289,900 videos from 73 signers covering 3,215 Auslan glosses, with 22 videos per gloss providing strong class balance. Only the front-facing RealSense camera was used, reflecting real-world single-camera deployment constraints. Evaluation used the Test-STU subset, which matches the background conditions of the training data, isolating model behaviour from environmental variation.

Custom BSL Dataset. No large-scale publicly available dataset exists for isolated BSL recognition [15]. A custom dataset was therefore constructed by scraping and manually segmenting publicly available BSL educational videos from YouTube and dedicated BSL teaching sites, in accordance with approved ethical guidelines. This

process introduced naturalistic variability in backgrounds, lighting, and signer style. Each sign was independently verified for accuracy against SignBSL. The resulting dataset is class-imbalanced with a modal class size of 12 videos and some glosses having as few as six training samples.

4.2 Vocabulary Selection

Experiments were conducted across two vocabulary stages applied to both Auslan and BSL. Stage 1 comprised alphanumeric signs (letters A–Z and numbers 0–10 if available). Stage 2 expanded to a 116-class Auslan and 85-class BSL vocabulary drawn from an adapted Swadesh list [20], a linguistically motivated selection of high-frequency vocabulary. Dataset statistics are summarised in Table 1.

Table 1: Dataset statistics for Auslan and BSL across both vocabulary stages (train/val/test videos).

Dataset	Stage	Classes	Train	Val/Test
Auslan	1	35	420	70
Auslan	2	116	1,392	170
BSL	1	36	288	72
BSL	2	85	622	170

4.3 Evaluation Framework

Quantitative evaluation used per-class accuracy, precision, recall, and macro-averaged F1-score, with confusion matrices to analyse misclassification patterns. For BSL, scratch-trained models were compared against transfer-learned models to isolate the contribution of cross-lingual transfer.

To complement test metrics, a qualitative and statistical attention analysis framework was developed. Attention weights from the final ViT transformer layer were extracted and overlaid on representative frames as heatmaps. Four regions of interest (ROIs), Person, Face, Arms, and Hands, were detected automatically using Mediapipe pose and face estimation models, with extrapolation applied to handle occlusion and motion blur. Pre- and post-fine-tuning attention concentration was compared using Wilcoxon signed-rank tests with effect sizes quantified by Cohen’s d . Spearman’s rank correlation was computed between per-ROI attention concentration and classification accuracy to assess whether attention on linguistically relevant regions predicts correct classification. A Bonferroni correction was applied across the four ROIs ($\alpha = 0.0125$).

5 RESULTS

5.1 Auslan Stage 1 and Stage 2

The full architecture achieved 71.42% validation accuracy on Stage 1 (35-class alphanumeric vocabulary). Scaling to Stage 2 (116-class Swadesh vocabulary) yielded 77.59% validation accuracy, despite a $3.3\times$ increase in class count at the same per-class training density. On unseen testing, 55.71% test accuracy on Stage 1 was achieved. Scaling to Stage 2 improved to 78.45% test accuracy with an F1 of 76.12% (Table 2).

The improvement from Stage 1 to Stage 2 despite the larger vocabulary is attributable to the nature of the signs. Alphanumeric finger-spelling involves subtle static handshapes with minimal inter-class variation, whereas Swadesh lexical signs involve more varied arm

Table 2: Auslan test performance across vocabulary stages.

Stage	Acc (%)	Prec (%)	Rec (%)	F1 (%)
1	55.71	46.86	55.71	48.35
2	78.45	79.01	78.45	76.12

movements and body configurations that provide clearer decision boundaries for the temporal modelling component.

Attention Analysis. Table 3 reports Wilcoxon and Spearman results for both Auslan stages. Fine-tuning produced statistically significant increases in attention concentration across all four ROIs at both stages, with effect sizes growing substantially from Stage 1 to Stage 2. The Face ROI reached significance at both stages but with a negligible effect size, consistent with the limited role of non-manual features in isolated lexical recognition.

Table 3: Attention analysis for Auslan models. All significant results survive Bonferroni correction ($\alpha = 0.0125$).

Stage	ROI	d	Wilcoxon	ρ	Corr.
1	Arms	1.46	Yes	0.487	Yes
1	Face	0.36	Yes	0.145	No
1	Hands	1.19	Yes	0.544	Yes
1	Person	1.87	Yes	0.432	Yes
2	Arms	2.02	Yes	0.660	Yes
2	Face	0.06	Yes	0.100	No
2	Hands	1.52	Yes	0.650	Yes
2	Person	2.84	Yes	0.650	Yes

Spearman correlations were significant for Hands, Arms, and Person at both stages ($\rho = 0.43$ – 0.66), strengthening considerably from Stage 1 to Stage 2. The Arms ROI showed the largest increase ($\rho = 0.487 \rightarrow 0.660$), indicating that temporal arm movement information became increasingly predictive of correct classification as vocabulary complexity grew. Face attention showed no significant correlation at either stage.

Error Analysis. Confusion matrix analysis on the Stage 1 test split revealed consistent misclassification between visually similar handshape pairs. K was frequently predicted as D, R as U, and Three as Four, all cases where signs share near-identical spatial configurations distinguishable primarily by subtle wrist orientation or minor finger extension. These confusions are consistent with the lower Stage 1 accuracy relative to Stage 2. Pairs such as N/E and Y/T showed less visually motivated confusions, suggesting the model may require dedicated hand keypoint features to resolve fine-grained handshape distinctions.

5.2 BSL Cross-Lingual Transfer

Table 4 reports test performance for all four BSL configurations. Transfer learning from Auslan substantially outperformed training from scratch in both stages.

Table 4: BSL test performance: scratch training vs. Auslan-to-BSL transfer.

Stage	Type	Acc (%)	Prec (%)	Rec (%)	F1 (%)
1	Scratch	11.11	7.54	11.11	8.32
1	Transfer	25.71	22.08	25.71	21.54
2	Scratch	28.24	25.88	28.24	25.14
2	Transfer	38.82	38.98	38.82	35.79

At Stage 1, transfer learning doubled test accuracy relative to the scratch baseline (11.11% to 25.71%, a 14.6 percentage point gain). At Stage 2, the gain was 10.6 percentage points (28.24% to 38.82%). Table 5 reports the corresponding attention statistics.

Table 5: Attention analysis for BSL models. Cohen’s d reports the pre/post fine-tuning effect size; Spearman ρ reports correlation with classification accuracy. Significance after Bonferroni correction ($\alpha = 0.0125$).

Stage	Type	ROI	d	Wilcoxon	ρ	Corr.
1	Scratch	Arms	0.00	No	−0.16	No
1	Scratch	Hands	0.00	No	−0.13	No
1	Transfer	Arms	−0.10	Yes	0.03	No
1	Transfer	Hands	0.36	Yes	0.15	No
1	Transfer	Person	0.24	Yes	0.10	No
2	Scratch	Arms	0.51	Yes	0.04	No
2	Scratch	Hands	0.60	Yes	0.02	No
2	Scratch	Person	0.79	Yes	0.03	No
2	Transfer	Arms	0.10	Yes	−0.06	No
2	Transfer	Hands	0.62	Yes	0.11	No
2	Transfer	Person	0.35	Yes	−0.02	No

The Stage 1 scratch model showed no measurable change in attention concentration following fine-tuning (all $d = 0.0$), indicating that the unfrozen backbone received no effective gradient signal under extreme data scarcity. The Stage 1 transfer model showed significant shifts in Hands and Person attention, confirming that the Auslan initialisation provided a stable enough foundation for spatial learning to proceed. Figure 2 illustrates this contrast: the transfer model develops clear hand-focused attention after fine-tuning, while the scratch model produces no comparable shift.



(a) Scratch Model

Figure 2: Temporal attention visualisation for BSL Stage 2 Scratch vs Transfer after training.

Spearman correlations between attention concentration and classification accuracy were not statistically significant in any BSL configuration, in marked contrast to the Auslan results. Auslan models were trained and evaluated on the same signer pool with consistent studio backgrounds, enabling stable correspondences between spatial attention and linguistic content. BSL models faced naturalistic recording variability combined with as few as eight training samples per class, insufficient to establish reliable attention-accuracy relationships even when spatial attention did shift toward relevant regions.



(b) Transfer Model

Figure 2: Temporal attention visualisation for BSL Stage 2 Scratch vs Transfer after training (continued).

6 DISCUSSION

6.1 Scalability Across Vocabulary Complexity

The improvement from Stage 1 to Stage 2 despite a $3.3\times$ class count increase reflects a fundamental property of sign language vocabulary rather than a limitation of the Stage 1 task. Alphanumeric fingerspelling compresses 35 classes into a highly confusable handshape space in which many letters differ only in subtle wrist rotation, finger extension, or brief movement. The Swadesh lexical signs, by contrast, involve full arm trajectories, body locations, and temporal dynamics that create well-separated decision boundaries.

The Spearman correlation data quantifies this shift precisely. The Arms ROI correlation with classification accuracy increased from $\rho = 0.487$ at Stage 1 to $\rho = 0.660$ at Stage 2, indicating that the Bi-LSTM’s temporal modelling of arm movement became substantially more predictive of correct classification as vocabulary complexity grew. This is precisely the computational advantage of the hybrid architecture over static image approaches: arm movement over time is the modality that separates lexical signs most clearly, and it is only captured by the temporal component. The Hands correlation also strengthened ($\rho = 0.544 \rightarrow 0.650$), while the Face ROI remained non-significant at both stages. These results suggest the hybrid architecture’s value increases with vocabulary complexity, a practically important property for any system intended to support classroom communication rather than laboratory fingerspelling.

6.2 Cross-Lingual Transfer from Auslan to BSL

The Stage 1 scratch model showed no measurable change in spatial attention following fine-tuning (all $d = 0.0$), indicating that the unfrozen backbone received insufficient gradient signal to update its representations at all. Training from scratch on 288 samples across 36 classes did not provide enough signal to shift a randomly initialised backbone. The transfer model, starting from Auslan-adapted representations, had enough spatial prior to begin learning under the same data conditions. This is consistent with Mocialov et al.’s finding [15] that cross-lingual transfer produces better initial performance even when source and target languages are structurally distinct.

At Stage 2, the transfer benefit narrowed, with scratch and transfer models converging to 25.14% and 35.79% F1 respectively. This convergence pattern is consistent with data scarcity becoming the dominant constraint at higher vocabulary complexity. With 85 classes and an average of seven to eight training samples per class, neither initialisation strategy can overcome the fundamental limitation of insufficient labelled data. The architecture and pretraining are not the bottleneck at Stage 2: the BSL dataset is.

6.3 Attention Analysis as a Reusable Evaluation Framework

The attention analysis framework developed here provides statistically grounded evidence of whether a model is learning linguistically motivated representations, which is a methodological contribution applicable to future spatiotemporal SLR research. The divergence between Auslan and BSL Spearman results illustrates the framework’s diagnostic value. Both model families showed significant attention shifts toward hands and arms following fine-tuning. Yet only the Auslan models showed significant correlations between that attention and correct classification. This distinction reveals that attending to the right region and reliably using that region for classification are separable phenomena. The BSL models learned to look in approximately the right places but lacked sufficient data to convert spatial attention into reliable discriminative signal. A framework measuring only attention maps or only classification accuracy would have missed this distinction entirely.

6.4 Educational Implications and Limitations

The results have direct implications for the development of assistive technology in educational settings. Over 90% of Deaf children are born to hearing, non-signing families, placing them at acute risk of language deprivation [12]. An SLR system capable of providing real-time signing feedback in classroom and home environments could meaningfully reduce this risk, supporting early language exposure and independent vocabulary acquisition. The Swadesh vocabulary evaluated in Stage 2 was selected precisely because it reflects high-frequency communicative vocabulary relevant to everyday educational interaction rather than laboratory sign sets. The 78.45% accuracy achieved on this vocabulary, under low-resource constraints with approximately 12 training samples per class, demonstrates that the architecture is viable for practical deployment in educational assistive tools without requiring large institutional data collection efforts.

The most significant limitation is BSL dataset scale. With a small number of training samples per class drawn from the same small pool of signers, models are vulnerable to learning signer-specific features rather than linguistically general ones. True signer-independent evaluation would be an essential step before deployment. Covariate shift between clean Auslan studio recordings and naturalistic BSL educational videos also limits transferability, and likely accounts for a meaningful fraction of the performance gap between Auslan and BSL results. Future work should prioritise dedicated BSL data collection, few-shot learning approaches [15], and extension toward real-time classroom feedback tools that could provide DHH learners and their families with accessible, low-cost signing support.

7 CONCLUSION

This paper presented a hybrid ViT-LSTM architecture for low-resource word-level sign language recognition, motivated by the need

for accessible assistive tools for DHH learners in educational environments. The system was evaluated across two vocabulary stages on Auslan and via cross-lingual transfer to BSL. Three findings emerged. First, the architecture scales effectively with vocabulary complexity: Stage 2 achieved 78.45% test accuracy on 116 classes and outperformed Stage 1 on 35 classes, attributed to the greater temporal distinctiveness of lexical signs over fingerspelling. Second, cross-lingual transfer from Auslan to BSL provides meaningful gains under extreme data scarcity, doubling test accuracy at Stage 1 despite structural differences between the two languages’ fingerspelling systems. Third, the attention analysis framework developed here provides interpretable and statistically grounded evidence of linguistically motivated spatial learning, with Spearman correlations confirming that Hands and Arms attention predicts correct classification in Auslan models while simultaneously revealing that data scarcity prevents this relationship from forming in BSL models.

Future work should prioritise larger BSL dataset construction, signer-independent evaluation, and incorporation of explicit hand keypoint features. Extension to continuous SLR and classroom learner feedback tools would realise the educational motivation of this work, contributing toward equitable AI-supported education for DHH communities.

REFERENCES

- [1] ‘Lexical Comparison of Signs From American, Australian, British, and New Zealand Sign Languages: David McKee and Graeme Kennedy’, in *The Signs of Language Revisited*, Psychology Press, (2000).
- [2] Ben Ambridge, ‘A Computational Simulation of Children’s Language Acquisition (Crazy New Idea)’, *OASICS, Volume 93, LDK 2021*, 93, 4:1–4:3, (2021). <https://drops.dagstuhl.de/entities/document/10.4230/OASICS.LDK.2021.4>.
- [3] Huilong Chen, Jianjiang Cui, Yucheng Zhang, and Yonglin Zhang, ‘ViT and Bi-LSTM for Micro-Expressions Recognition’, in *2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pp. 946–951, (September 2022). <https://ieeexplore.ieee.org/document/9927522>.
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. <http://arxiv.org/abs/2010.11929>, June 2021.
- [5] Faye Ginsburg, ‘Disability in the Digital Age’, in *Digital Anthropology*, Routledge, (2012).
- [6] Sepp Hochreiter and Jürgen Schmidhuber, ‘Long Short-Term Memory’, *Neural Computation*, 9(8), 1735–1780, (November 1997).
- [7] Jeremy Howard and Sebastian Ruder. Universal Language Model Fine-tuning for Text Classification. <http://arxiv.org/abs/1801.06146>, May 2018.
- [8] Jiayu Huang and Varin Chouvatut, ‘Video-Based Sign Language Recognition via ResNet and LSTM Network’, *Journal of Imaging*, 10(6), 149, (June 2024). <https://www.mdpi.com/2313-433X/10/6/149>.
- [9] Tisha Jantzen, ‘The Power of Sign: Enhancing Oral Communication with Young Children with Typical Hearing’, *Research Papers*, (January 2011). https://opensiuc.lib.siu.edu/gs_rp/171.
- [10] Zhenni Li, Yujie Li, Benying Tan, Shuxue Ding, and Shengli Xie, ‘Structured Sparse Coding With the Group Log-regularizer for Key Frame Extraction’, *IEEE/CAA Journal of Automatica Sinica*, 9(10), 1818–1830, (October 2022). <https://ieeexplore.ieee.org/document/9774986>.
- [11] Zeyu Liang, Huailin Li, and Jianping Chai, ‘Sign Language Translation: A Survey of Approaches and Techniques’, *Electronics*, 12(12), 2678, (June 2023). <https://www.mdpi.com/2079-9292/12/12/2678>.

- [12] Diane Lillo-Martin and Jonathan Henner, ‘Acquisition of Sign Languages’, *Annual Review of Linguistics*, 7(1), 395–419, (January 2021). <https://pmc.ncbi.nlm.nih.gov/articles/PMC8570554/>.
- [13] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. <http://arxiv.org/abs/1711.05101>, January 2019.
- [14] M. Madhwarasan and Partha Pratim Roy. A Comprehensive Review of Sign Language Recognition: Different Types, Modalities, and Datasets. <http://arxiv.org/abs/2204.03328>, April 2022.
- [15] Boris Mocialov, Graham Turner, and Helen Hastie. Transfer Learning for British Sign Language Modelling. <http://arxiv.org/abs/2006.02144>, June 2020.
- [16] Xin Shen, Heming Du, Hongwei Sheng, Shuyun Wang, Hui Chen, Huiqiang Chen, Zhuojie Wu, Xiaobiao Du, Jiaying Ying, Ruihan Lu, Qingzheng Xu, and Xin Yu. MM-WLAuslan: Multi-View Multi-Modal Word-Level Australian Sign Language Recognition Dataset. <http://arxiv.org/abs/2410.19488>, October 2024.
- [17] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. <http://arxiv.org/abs/2106.10270>, June 2022.
- [18] B Sundar and T Bagyammal, ‘American Sign Language Recognition for Alphabets Using MediaPipe and LSTM’, *Procedia Computer Science*, **215**, 642–651, (2022). <https://www.sciencedirect.com/science/article/pii/S1877050922021378>.
- [19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. <http://arxiv.org/abs/1512.00567>, December 2015.
- [20] James Woodward, ‘Sign languages and sign language families in Thailand and Viet Nam’, *The Signs of Language Revisited: An Anthology in Honor of Ursula Bellugi and Edward Klima*, (January 2000).
- [21] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? <http://arxiv.org/abs/1411.1792>, November 2014.

Democratising Personalised STEAM Education: A Generative AI-Driven Intelligent Tutoring System Aligned with Standardised Pedagogical Frameworks in Mexico

Alma Jazmín Castillo Rodríguez¹ and Felipe De la Peña²

ABSTRACT.

This paper presents the design, and implementation, of an Adaptive STEAM Virtual Tutor — an Intelligent Tutoring System (ITS) built on Google Colab and powered by the Gemini Flash generative model via the Google Gen AI Python SDK. The system automates a three-stage pedagogical pipeline: (1) diagnostic assessment delivered through Google Forms and recorded in Google Sheets, (2) algorithmic performance classification into three levels — Bajo ($\leq 50\%$), Normal ($60\%–75\%$), and Sobresaliente ($80\%–100\%$) — using a custom Python function, and (3) personalised activity generation by Gemini Flash, which produces level-differentiated reinforcement exercises and teacher recommendations via a structured prompt.

The system's evaluation consisted of an initial pilot in Mathematics and Complexity Sciences at CBTIS 131 (a public federal upper secondary school- equivalent to Sixth Form in the United Kingdom) with a cohort of fewer than 20 students over two weeks. Results validated against independent CENEVAL-aligned post-tests show a 20% average group improvement (6.5/10 to 8.5/10), individual gains from 50% to 100% in the primary case study, and a reduction in teacher profiling time from approximately one week to under five minutes. The system is distinguished by its zero-cost deployment on pre-existing Google Workspace infrastructure, its alignment with both the CENEVAL framework and the DGETI/SEP Marco Curricular Común, and its affective pedagogical design targeting math anxiety reduction

Keywords: Intelligent Tutoring Systems, Generative AI, Gemini Flash, Google Colab, STEAM Education, CENEVAL, Mexico, Adaptive Learning, Bachillerato, MCC DGETI.

1 INTRODUCTION

The global education landscape faces a persistent and widening challenge: ensuring that students develop genuine proficiency in STEAM disciplines—Science, Technology, Engineering, Arts, and Mathematics—at a time when classroom instruction increasingly struggles to meet individual learners where they are. As Anderson, Boyle, and Reiser [1] observed four decades ago, Intelligent Tutoring Systems are "programs that simulate an understanding of the domain they teach and can respond specifically to the student's problem-solving strategies." That foundational insight has lost none of its relevance; if anything, the emergence of generative artificial intelligence has made it more urgent and more achievable than ever before.

In Mexico's upper secondary education—*Educación Media Superior*—these pressures are especially acute. Students at public technical institutions such as *Centro de Bachillerato Tecnológico*

Industrial y de Servicios Núm. 131 (CBTIS 131) regularly encounter conceptual thresholds—self-similarity in fractal geometry, non-linear dynamics in Complexity Sciences—that a single 50-minute session, shared among an entire heterogeneous cohort, cannot adequately address. National curricular frameworks, from the *Marco Curricular Común* (MCC) of the DGETI/SEP system to the assessment guidelines of CENEVAL, demand competency-based learning and higher-order thinking [2,3], yet the resources available to teachers in these settings remain severely constrained.

Research spanning several decades offers a compelling response to this gap. VanLehn [4] demonstrated that well-designed computer-based tutors can approach the effectiveness of one-on-one human instruction in specific educational contexts, while Graesser et al. [5] showed, through AutoTutor, the particular power of natural language dialogue in fostering deeper conceptual understanding. Meanwhile, Woolf [6] emphasized the importance of adaptive learning strategies that respond dynamically to each student's performance, pace, and cognitive needs. Together, these findings point toward a clear design agenda: personalized, conversational, and adaptively responsive tutoring at scale.

Recent advances in large language models have brought that agenda within reach for everyday educational contexts. Google's Gemini Flash model, accessible through open programming interfaces, offers sophisticated natural language understanding and real-time content generation [7]—capabilities that align closely with Krashen's [8] Input Hypothesis, which stresses the importance of comprehensible, context-sensitive interaction for meaningful learning. By embedding such a model within accessible, zero-cost infrastructure, it becomes possible to build an intelligent tutor that is not only pedagogically sound but also practically deployable in under-resourced schools.

This article presents the design and implementation of the *Adaptive STEAM Virtual Tutor*, an Intelligent Tutoring System developed on Google Colab and powered by the Gemini Flash generative model via the Google Gen AI Python SDK. Piloted at CBTIS 131 in San Luis Potosí, the system integrates adaptive questioning, formative feedback, automated assessment generation, while remaining fully aligned with the CENEVAL national evaluation framework and the MCC competency standards of the DGETI/SEP system. The study describes the system's architecture, its pedagogical foundations, and the implementation process, demonstrating how state-of-the-art generative AI can serve as the engine of an effective, scalable, and curriculum-aware educational tool for students in *Educación Media Super*

¹ CBTIS 131, DGETI/SEP, San Luis Potosí, México — almajazmin.castillo.cb131@dgeti.sems.gob.mx

² AT&T USA — fd7296@att.com

2 RELATED WORK AND CONTRIBUTIONS

Contemporary ITS solutions span three paradigms as shown in figure 1: constraint-based systems (Carnegie Learning MATHia) use Bayesian Knowledge Tracing (BKT) and expert-encoded rule sets to deliver deterministic, step-level feedback; over two decades of empirical evidence confirm learning gains comparable to one-on-one human tutoring [9, 10], though their construction requires extensive domain engineering by subject-matter experts [11].

Data-driven systems such as ASSISTments infer student knowledge states from large interaction logs rather than hand-coded rules [12]; advances in deep learning — from Deep Knowledge Tracing [13] to Transformer-based architectures [14] — have substantially improved predictive accuracy, while the platform's open dataset infrastructure has established it as a benchmark for educational data mining research.

LLM-based tutors such as Khan Academy's Khanmigo and GPT-4-powered systems represent the most recent paradigm, offering natural-language dialogue and broad domain coverage; a systematic review of experimental studies reports improvements in academic performance, affective-motivational states, and higher-order thinking [15], yet hallucination — the generation of plausible but factually incorrect feedback — remains a critical reliability concern documented across mathematics and science domains [16, 17].

Figure 1. Intelligent Tutoring Sysyem paradigms illustrated with everyday analogies



As illustrated in Figure 1, the three paradigms differ not only in their underlying technical architecture but also in where they place the burden of pedagogical knowledge: constraint-based systems locate it in the expert who encodes the rules; data-driven systems distribute it across the historical behaviour of millions of students; and LLM-based systems embed it implicitly within the parameters of a pre-trained language model. This distinction has direct consequences for scalability, reliability, and contextual validity. While each paradigm offers a distinct combination of strengths — diagnostic precision, data-driven adaptability, and conversational fluency, respectively — none addresses simultaneously the cost

barriers of under-resourced public education systems, the absence of alignment to non-Anglo-American curricula, and the affective dimensions of learning in high-stakes secondary contexts. The proposed system is designed precisely at this intersection: it inherits the conversational capability of LLM-based tutors while constraining generative output to formally specified learning objectives drawn from the CENEVAL item criteria and the DGETI/SEP MCCEMS competency framework, functioning as a structural hallucination-mitigation strategy absent from all three existing paradigms as currently deployed.

Against this backdrop, our system is differentiated along three axes:

- **Zero-cost deployment:** The entire stack — Google Forms, Google Sheets, Google Colab, and the Gemini Flash API — is available at no cost to Mexican public schools already operating within the Google Workspace for Education ecosystem, unlike licensed ITS platforms.
- **Mexican curricular alignment:** Existing ITS solutions predominantly target Anglo-American curricula. Our system is explicitly designed under CENEVAL criteria for formally structured test items and the DGETI/SEP Marco Curricular Común (MCCEMS), providing construct validity within the Mexican formal assessment context through a `definer_nivel()` function that maps each generated activity to a specific competency level prior to delivery. Critically, this curricular anchoring also serves as a hallucination-mitigation strategy: by constraining generative output to formally specified learning objectives and competency levels, the system reduces the risk of off-domain or pedagogically unsound content generation [16].
- **Affective pedagogy:** The system's prompt engineering incorporates an empathetic teacher persona and real-world analogies to reduce the affective filter [1] — a documented barrier to STEAM learning in high-stakes Mexican secondary contexts.

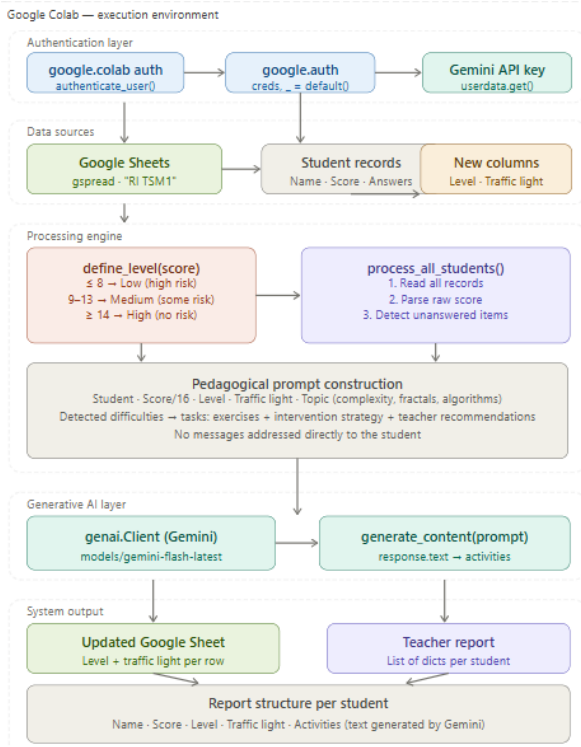
3 SYSTEM ARCHITECTURE AND TECHNICAL IMPLEMENTATION

The system is implemented as a Google Colab notebook using Python. The full technical stack comprises: the `google-generativeai` and `gsread` Python libraries, the Google Colab authentication module, the Gemini Flash generative model (`models/gemini-flash-latest`), and Google Sheets as the persistent data layer.

As shown in Figure 2, the pipeline consists of five layers: (1) an *authentication layer* using `google.colab.auth` and `google.auth` to obtain credentials for Google Sheets and the Gemini API; (2) a *data sources layer* that reads student records (name, score, answers) from the Google Sheets spreadsheet "RI TSM1" via `gsread`; (3) a *processing engine* that applies the `define_level(score)` function to classify students into three risk levels (Low: ≥ 8 , Medium: 9–13, High: ≥ 14) and the `process_all_students()` function to parse scores and detect unanswered items; (4) a *pedagogical prompt construction layer* that assembles personalized prompts encoding student score, risk level, topic, detected difficulties, and teacher recommendations; and (5) a *generative AI layer* using `genai.Client` (Gemini Flash) and `generate_content(prompt)` to produce individualized activity sets. System outputs include an updated Google Sheet with level and traffic-light indicators per row, and a teacher report as a

list of dictionaries per student, structured by name, score, level, traffic light, and AI-generated activities.

Figure 2. SYSTEM ARCHITECTURE



3.1 Authentication Layer

The notebook begins by installing and importing the required libraries and authenticating against Google services:

Figure 3. Python authentication setup: library imports, Gemini API key retrieval from Colab secrets, and OAuth2-bases Google Sheets authorization

```
!pip install --upgrade -q google-generativeai gspread

from google.colab import userdata, auth
from google import genai
import gspread
from google.auth import default

API_KEY = userdata.get('GEMINI_API_KEY')
client = genai.Client(api_key=API_KEY)

auth.authenticate_user()
creds, _ = default()
gc = gspread.authorize(creds)
```

Figure 3 shows the authentication steps. The Gemini API key is stored as a Colab secret (userdata), avoiding hardcoded credentials. Google Sheets authentication uses OAuth2 via the Colab built-in flow, ensuring institutional data remains within the Google Workspace security perimeter.

3.2 Data Source: Google Sheets Integration

Figure 4. Work flow for reading evaluation records from Google Forms into the system

```
SPREADSHEET_NAME = "ED TSM1"
sheet = gc.open(SPREADSHEET_NAME).sheet1
data = sheet.get_all_records()
ultimo_estudiante = data[-1] # most recent submission
```

As shown in Figure 4, student diagnostic data is read directly from a Google Sheets spreadsheet named 'ED TSM1', which collects responses from a Google Forms diagnostic assessment. The system accesses the first sheet and reads all records as a list of dictionaries.

Figure 5. Google Sheets data ingestion: reading students records from the spreadsheet and parsing the diagnostic score

```
valor_sheet = ultimo_estudiante.get('Puntuación (puntuación máxima 8/10)', 0)
puntuacion = int(str(valor_sheet).split()[0])
```

As seen in Figure 5, the diagnostic score is extracted from the column 'Puntuación (puntuación máxima 8/10)', with robust string parsing to handle Google Sheets' varied numeric formats.

3.3 Algorithmic Performance Classification

A custom Python function classifies each student's score (out of 8) into one of three performance levels ("Bajo" "Normal" "Sobresaliente") as shown in Figure 6.

Figure 6. Python implementation of definir_nivel

```
def definir_nivel(p):
    if p <= 4:
        return "Bajo" # Basic: guided exercises
    elif 5 <= p <= 6:
        return "Normal" # Intermediate: applied practice
    else:
        return "Sobresaliente" # Advanced: deepening tasks
```

The level determination algorithm — Bajo (≤4/8, 0–50%), Normal (5–6/8, 62.5–75%), Sobresaliente (7–8/8, 87.5–100%) — maps directly to differentiated instructional strategies, ensuring the AI-generated content is calibrated to the student's specific learning needs rather than delivering uniform content.

For students at a low achievement level, the system generates lessons comprising guided exercises and explanations tailored to the student's content. For intermediate-level students, an application exercises are provided, whereas for high-achieving students, exercises are generated, requiring specific research or more complex problem solving.

3.4 Personalised Lesson Generation via Gemini Flash

The core of the system is a structured prompt passed to the Gemini Flash model (models/gemini-flash-latest), which generates level-differentiated reinforcement activities and teacher recommendations. The prompt shown in figure 7 incorporates the student's name, diagnostic score, performance level, and explicit pedagogical directives for each level

The prompt instructs Gemini Flash to adopt the persona of an expert teacher in Mexican upper secondary education, generating content appropriate to the student's diagnosed performance level.

Figure 7. Structured pedagogical prompts, that produce level-diferenriated reinforcemente activitties and teacher recommendations,

```
prompt = f"""
Eres un docente experto en educación media superior.

Alumno: {nombre}
Puntuación: {puntuacion}/8
Nivel de desempeño: {nivel_actual}

Diseña actividades de refuerzo acordes al nivel:
- Bajo: ejercicios guiados
- Normal: aplicación
- Sobresaliente: profundización

Incluye recomendaciones para el docente.
"""

response = client.models.generate_content(
    model="models/gemini-flash-latest",
    contents=prompt
)
```

The explicit level-strategy mapping (guided exercises for Bajo, applied practice for Normal, deepening tasks for Sobresaliente) ensures the AI's output is pedagogically differentiated rather than generic. The inclusion of teacher recommendations within the same output serves the dual function of supporting the student and reducing the teacher's planning burden.

3.5 System Output

The function returns shown in figure 8 a structured output combining the student's profile data and the AI-generated activities:

Figure 8. Structured return value of the tutoring piprlinr, formatted for direct sharing with both students and teacher?

```
return f"""
Alumno: {nombre}
Puntuación: {puntuacion}/8
Nivel: {nivel_actual}

--- ACTIVIDADES SUGERIDAS ---
{response.text}
"""
```

This output is human-readable and designed to be shared directly with both the student (personalised reinforcement plan) and the

teacher (recommendations for follow-up instruction), making the system immediately actionable within the classroom workflow.

4 METHODOLOGY AND PILOT IMPLEMENTATION

4.1 Pilot Design

The pilot was conducted at CBTIS 131, San Luis Potosí, México, over two weeks with a cohort of fewer than 20 students enrolled in Mathematics and Complexity Sciences (Pensamiento matemático / sistemas complejos) at upper secondary level. The study followed a pre-post intervention design with three measurement points: (1) baseline diagnostic via Google Forms (scores recorded in Google Sheets 'ED TSM1'), (2) AI-generated personalised activity delivery, and (3) independent post-intervention assessment using CENEVAL-aligned instruments.

A critical methodological decision was the strict separation between the instructional engine (Gemini Flash-generated activities) and the evaluative framework (independent CENEVAL-aligned post-tests). This separation ensures that measured learning gains reflect genuine acquisition rather than familiarity with the AI's output format.

4.2 Case Study: Wendy

To illustrate the system's operation, we present the case of one participant — anonymised as *Wendy* — whose Google Forms diagnostic score of 4/8 classified her as *Bajo*, triggering the guided-exercise pedagogical strategy. The system identified a conceptual gap between symbolic procedural fluency and applied understanding of non-linear systems.

The Gemini Flash-generated output for Wendy's level adopted an empathetic teacher persona, using urban traffic flow as an analogy for non-linear system behaviour — a real-world contextualisation strategy targeting the affective filter. On the independent CENEVAL-aligned post-test, Wendy achieved 100% (16/16), compared to her baseline of 50% (4/8).

5 ANALYSIS OF RESULTS

Table 1. SUMMARY PILOT RESULTS

Evaluation Metric	Baseline	Post-Intervention	Net Variance & Pedagogical Impact
Avg. group score — Mathematics	6.5 / 10	8.5 / 10	+20% group improvement
Avg. group score — Complexity Sciences	6.2 / 10	8.1 / 10	+19% relative gain
Case study — Wendy (Bajo → Sobresaliente)	4 / 8 (50%)	16 / 16 (100%)	Performance recovery (+50%)
Systemic Efficiency: Teacher profiling time	~1 week (manual)	~5 min (automated)	>99% reduction in profiling time
Early At-Risk Identification (Bajo Level)	0% (reactive)	100% (predictive)	Automated diagnostic triggering

5.1 Controlled Pedagogical Performance and Withdrawal Effects

Average group scores in Mathematics improved from 6.5/10 to 8.5/10 (20% gain) on CENEVAL-aligned instruments; in Complexity Sciences from 6.2/10 to 8.1/10 (19% relative gain). The primary case study demonstrates maximal performance recovery in a single intervention cycle, validating the level-classification-to-prompt-strategy mapping implemented in the `definir_nivel()` function.

5.2 Systemic Efficiency

The automated pipeline — from Google Forms submission to Google Sheets ingestion to Gemini Flash generation — reduced student profiling time from approximately one week (manual assessment, error analysis, and differentiated lesson preparation) to under five minutes per student. This efficiency gain is structurally significant for under-resourced Mexican public institutions where one teacher may serve 40+ students per cohort.

5.3 Early At-Risk Identification

The `definir_nivel()` classification function identifies Bajo-level students ($\leq 4/8$) before disengagement, automatically triggering guided-exercise scaffolding protocols. This predictive capability — absent from standard classroom instruction — enables intervention at the point of first failure rather than after repeated poor performance.

6 DISCUSSION AND LIMITATIONS

The results confirm that effective ITS deployment does not require proprietary software. The system's reliance on the free Google Workspace ecosystem — universally available in Mexican public Bachillerato institutions — means it can be replicated by any teacher with a Google account, a Colab notebook, and a Gemini API key.

Several limitations must be acknowledged: (1) the cohort was small (<20 students) over a short period (2 weeks); (2) the single-institution setting limits generalisability; (3) absence of a control group prevents causal attribution of gains to the AI intervention alone; and (4) long-term retention was not measured. Additionally, the current implementation processes only the most recent Google Forms submission (`data[-1]`), which will require architectural modification for concurrent multi-student deployment in larger cohorts.

7 CONCLUSION AND FUTURE WORK

This paper has presented a fully implemented, zero-cost Intelligent Tutoring System for STEAM education in the Mexican upper secondary federal school, built on Google Colab, Google Sheets, and Gemini Flash. The system's three-stage automated pipeline — Forms-based diagnostic, Python classification, and Gemini-generated personalised activities — produces measurable learning gains validated against CENEVAL-aligned independent assessments, while reducing teacher profiling time by over 99%.

Future work will address the current limitations through: (1) parallel multi-student processing to replace the sequential `data[-1]` approach; (2) expansion to a multi-school cohort across Mexican states; (3) Within-Subject Validation and Cross-Partial Analysis; (4) longitudinal retention measurement; and (5) extension to additional DGETI curriculum subjects. The system offers a replicable blueprint for AI-augmented teaching that any upper

secondary federal school can implement at scale using existing infrastructure. With minimal adjustments to the underlying code and prompt design, this system can be effectively implemented across any educational level, offering a versatile tool for personalized learning environment.

ACKNOWLEDGEMENTS

The authors sincerely thank the educational authorities of San Luis Potosí: Secretaría de Educación de Gobierno del Estado (SEGE), Section 26 of the Sindicato Nacional de Trabajadores de la Educación (SNTE) and State Commissioner of DGETI. Juan Jose Mendoza López for the administrative arrangements and support that made attendance convention possible to present this work.

Special thanks are extended to the Principal of *Centro de Bachillerato Tecnológico Industrial y de Servicios No. 131 "Eugenio Garza Sada"* (CBTIS 131), Gerardo Alberto Nieto Flores, belonging to the DGETI system, for providing the necessary institutional access to conduct the pilot testing of this model.

Finally, the authors wish to express her sincere gratitude to Dr. María Margarita Rodríguez González for her invaluable pedagogical and methodological guidance, whose academic insights and support were fundamental to the structuring and development of this research.

REFERENCES

- [1] Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science*, 228(4698), 456–462.
- [2] SEP. (2022). Marco Curricular Común de la Educación Media Superior. DGETI. México.
- [3] CENEVAL. (2023). Guía para la construcción de reactivos de opción múltiple. México.
- [4] VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring
- [5] Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2005). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods*, 37(2), 180–192.
- [6] Woolf, B. P. (2010). Building Intelligent Interactive Tutors. Morgan Kaufmann.
- [7] Google. (2024). Gemini Flash API Documentation. Google DeepMind. <https://ai.google.dev/gemini-api/docs>
- [8] Krashen, S. D. (1982). Principles and Practice in Second Language Acquisition. Pergamon Press.
- [9] Holstein, K., et al. (2017). Intelligent tutoring systems: A meta-analysis. Cited in *Smart Learning Environments*, Springer.
- [10] Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2), 249–255.
- [11] Norberg, K., et al. (2024). Rewriting Math Word Problems with Large Language Models. *ERIC Grantee Submission*.
- [12] Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments Ecosystem. *International Journal of Artificial Intelligence in Education*, 24(4), 470–497.
- [13] Piech, C., et al. (2015). Deep Knowledge Tracing. *NeurIPS* 2015.
- [14] Abdelrahman, G., et al. (2023). Advanced Knowledge Tracing via Transformer-based framework. *JEDM (EDM Cup 2023)*.
- [15] Deng, X., et al. (2024). Systematic review on LLM interventions in education. Cited in arXiv:2507.02180.
- [16] Steinbach, M., Bhandari, S., Meyer, J., & Pardos, Z. A. (2025). When LLMs Hallucinate: Examining the Effects of

Erroneous Feedback in Math Tutoring Systems. *L@S 2025*,
Palermo.

- [17] Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLoS ONE*, 19(5), e0304013.