



Evgeny Chukharev, Iowa State University

Cognitive and Ethical Alignment of LLMs with Humans for Writing Research and Instruction

Large language models (LLMs) have transformed the study of writing. In linguistics, they catalyzed a shift from the generative grammar paradigm that dominated the latter half of the 20th century. Beyond their practical utility, LLMs provide strong empirical support for connectionist theories of human language processing, showing that complex linguistic behavior can emerge from statistical learning and distributed representations rather than relying solely on (innate) symbolic rules. At the same time, LLMs raise serious questions about alignment with human values, interpretability, and their impact on writing instruction and assessment. Constructing AI systems that simulate human linguistic behavior while aligning with human intentions, reasoning, and values offers both practical and research advantages.

This keynote presents two projects that illustrate how LLMs can be aligned with human cognition and ethical principles in writing research and instruction.

The first project leverages eye-tracking data, specifically writers' lookback fixations on text produced so far, to guide sentence completions in the emerging text. This approach operationalizes a long-standing hypothesis that writers look back at the text they have produced to support planning of what to say next. By conditioning LLM-based sentence completions on lookback fixation patterns, LLMs produce text that is more closely aligned with a writer's evolving intentions. This work provides empirical evidence for the cognitive function of lookback behavior and establishes a foundation for AI systems for writing support that operate in alignment with human cognitive processes.

The second project introduces a hybrid neurosymbolic AI framework for evaluating student argumentative writing. In this framework, LLM-driven inferences from source texts and student essays are constrained by symbolic reasoning that captures ethical norms, logical standards, and pedagogical criteria. By integrating the transparency and reliability of symbolic AI with the flexibility of LLMs in natural language understanding, this approach produces interpretable, robust evaluations of student writing that align with human ethical values.

Together, these projects demonstrate that aligning LLM behavior with human cognition and ethical principles can advance both the science of writing and instructional practice. By incorporating cognitive signals and symbolic constraints, AI systems can support and evaluate writing in ways that reflect human intentions, uphold reasoning standards, and promote responsible, interpretable applications of technology.