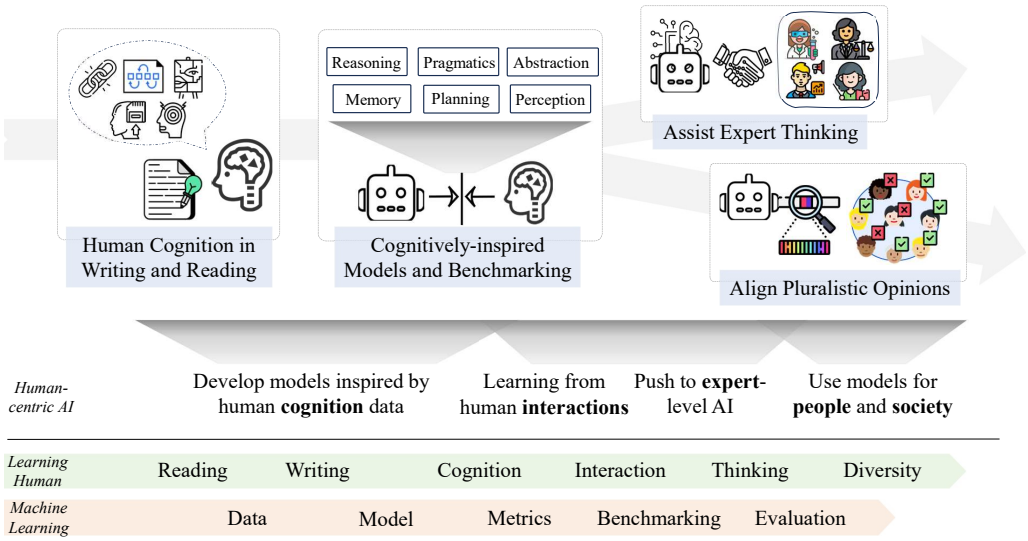
Cognitively-aligned AI for Experts, People, and Society

Dongyeop Kang
Last modified: August 2025

I aim to build human-centric language technologies, focusing on cognitively aligning human and machine thinking. My goal is to develop human-centric AI systems through the lens of *human cognition*, integrating insights from how people think, perceive, and collaborate into every stage of the machine learning lifecycle, from data creation to model design, evaluation, and human–AI interaction.



The central premise of my work is that AI should extend and enhance human cognition: supporting expert reasoning, fostering societal values, and accommodating diverse individual goals. While my long-term agenda spans multiple domains, I prioritize three tightly interconnected directions: (i) collecting and modeling human cognition data (§1), (ii) building cognitively-aligned AI models (§2) and evaluation frameworks (§3), and (iii) designing expert- (§4) and society-facing (§5) AI systems.

(§1) Understanding human cognition: Human cognition involves complex, multi-layered thinking processes. For example, writing is a non-linear and iterative cognitive process. By analyzing human writing data [16, 73, 49, 62], we can gain insights into these thought processes, which can then be used to improve the planning and reasoning capabilities of AI models [69, 15]. Similarly, human perception can be studied through reading behaviors, such as eye-tracking data or explicit perception annotations, offering clues on how people perceives and process information through reading [21, 10, 24]. Understanding both writing and reading behaviors enables us to enhance the cognitive capabilities of LLMs and develop AI assistants that better support human thinking.

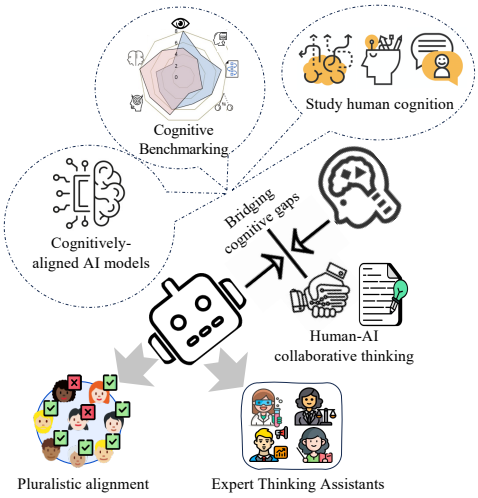
(§2) Cognitively-informed AI models: Cognitive alignment seeks to extend LLM capabilities to complex human behaviors, including neurosymbolic reasoning [40], task compositionality [69, 20], hierarchical planning [31, 36], abstraction, and social cognition [11]. We develop novel learning paradigms that either mimic human cognitive processes directly or incorporate cognitive data (e.g., writing, reading, expert reasoning, interaction) into training objectives. This includes self-supervised planning, structural alignment, multi-attribute alignment, and cognitively-efficient modeling strategies.

(§3) Cognitive benchmarking: Current LLM evaluation largely relies on shallow, monotonic tasks. We address this through two directions: (i) assessing both the potential [61, 77, 44, 45, 68, 43, 3] and risks [53, 7, 49, 62] of AI-generated data and AI-based evaluation, identifying issues such as cognitive bias, stylistic artifacts, and shallow synthesis, and (ii) building a large-scale cognitive assessment framework to evaluate and contrast human and machine cognition. The framework [12, 13] integrates rigorous cognitive science methodologies with scalable evaluation, in open collaboration with interdisciplinary researchers.

(§4) Expert-level AI: Assist Expert Thinking Process: In domains such as science and law, there is a

significant cognitive gap between human experts and current LLMs. We design interactive, domain-specific AI systems that facilitate productive collaboration, adapt to expert workflows, and provide cognitively-aligned assistance. For scientists, we have developed reading tools like *Semantic Reader* [24, 60] and *SciTalk* [67], as well as writing assistants that detect overloaded symbols, discourse-level inconsistencies, and promote iterative refinement [15, 73]. For legal professionals, we are collecting complex legal reasoning dataset *LawFlow* [8] and building legal assistants, that capture long-horizon planning and reasoning in legal tasks.

(§5) Pluralistic Alignment AI technologies must reflect diverse human perspectives to achieve societal alignment. We develop data-centric methods to detect, characterize, and augment underrepresented viewpoints [61, 72, 45, 42], and model-centric methods to encode pluralism at individual, group, and societal levels using distributional alignment, and societal value modeling [43, 22, 11]. This mitigates risks of monolithic outputs and supports socially inclusive AI.

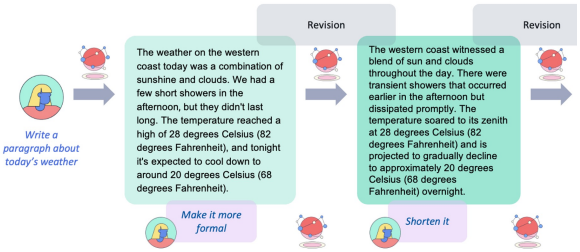


Together, these agendas form my long-term vision: *human-centric thinking partners* that perform complex, multi-step reasoning, reduce cognitive load, and reflect the diversity and values of human society. My research integrates linguistics, social sciences, and cognitive sciences, and is supported by industry and government partners such as Grammarly, NSF, CISCO, Sony, Accenture, and Open Philanthropy, with active collaborations with Amazon, AI2, Naver, and Google. I hold affiliations with cross-college labs at UMN and collaborate widely with faculty and students in computer science, law, psychology, education, journalism, design, and medicine. I also contribute to different synergetic activities through workshop organization: I co-organized the first "CtrlGen: Controllable Generative Modeling" workshop at NeurIPS 2021, and founded the "In2Writing: Intelligent and Interactive Writing Assistants" workshop series at ACL 2022, CHI 2023, and CHI 2024, fostering collaboration between ML, HCI, NLP, and professional writing communities. I also founded the "Pluralistic Alignment" workshop at NeurIPS 2024 to emphasize diversity in AI.

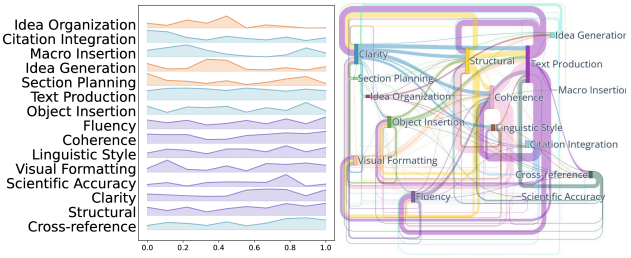
1 Collecting Human Cognition Data

Language serves as a window into human cognition. Writing and reading are not linear acts but complex, iterative processes that reveal how people plan, reason, and perceive information. For example, a well-written manuscript reflects a writer's long-term, multi-stage thinking, far beyond simple next-token prediction. Similarly, reading behaviors captured through eye-tracking or annotated perceptions provide insights into how individuals interpret, learn from, and engage with text. This first research pillar focuses on systematically collecting and analyzing human cognition data from both *production* (writing) and *perception* (reading) to better understand these processes and inform cognitively-aligned AI systems. We call this process as **cognitive-scaffolding** of LLMs learning from human cognition.

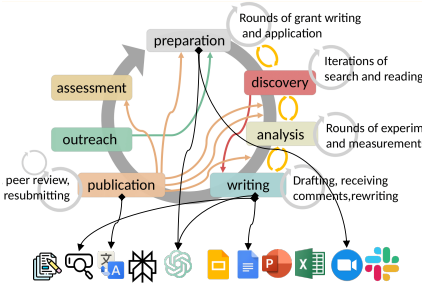
1.1 Writing and Workflow Data



Although large language models (LLMs) generate fluent text, human writing involves richer and more deliberate processes: drafting, revising, planning, proofreading, and synthesizing knowledge—often collaboratively. Domain-specific writing, such as scientific manuscripts, adds further challenges, requiring sustained practice, peer interaction, and integration of prior literature. We first examine the *iterative nature of text revision*. We have collected revision data across domains [16, 47] and built systems that emulate iterative refinement until quality stabilizes. Human–AI collaborative editing consistently outperforms either working alone [15, 69, 57]. Our studies show that iterative modeling of revision process by either AI only or human+AI collaboration has continuously shown better text quality, showing the **effective of test-time scaling of iterative revision or human-AI collaboration**. Also, guiding models with human revision intents (e.g., clarity, coherence) leads to better flow of revision traces and mimic human writing process, better designed supporting human writing.



(a) Scholarly Writing Intents over Time [73]

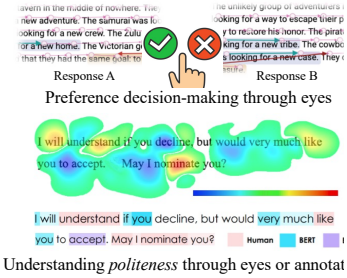


(b) Research Workflow with AI Tools

In *ScholarWrite* [54, 73], we extended prior work to capture the *entire* writing process[1]. This dataset provides a strong empirical basis for cognitively aligned writing assistants that model the full lifecycle of text production. Our vision is to build **a longitudinal dataset of complete research workflows**, tracing how researchers ideate, write, experiment, and collaborate across domains. Unlike datasets limited to individual actions or outputs, ours will record temporally grounded research sequences annotated with task intent, cognitive states, and AI interactions. The resulting corpus forms an **ontology of the scientific process**, enabling training of agents and interfaces that reflect real workflows. Ultimately, we aim to establish a longitudinal repository of cognitive workflows (science, law, education), allowing models to learn not isolated tasks but the temporal dynamics of human reasoning.

1.2 Reading Data

Whereas writing data reveals the generative processes of human thought, reading data captures the perceptual and evaluative side of cognition—how people interpret, prioritize, and react to information. We have collected explicit annotations of stylistic understanding [21] and used them to train LLMs that align with readers' stylistic judgments [23]. We have also gathered eye-tracking data to study how readers engage with stylistic cues and narrative content [10, 66],



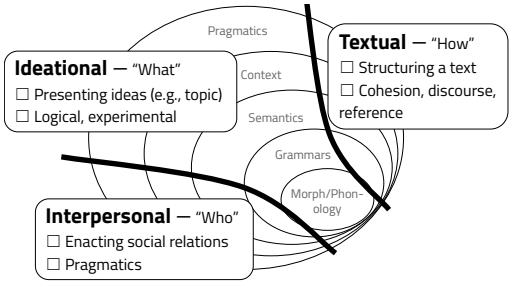
Understanding *politeness* through eyes or annotat

[1] 63,000 keystrokes of text edits with intention annotations from multi-month research papers

showing that gaze patterns provide richer cognitive signals than explicit annotations or model-based interpretations. [43]. More recently, we have collected human judgment data—such as preference rankings between texts—via both annotations and eye-tracking By collecting data that closely mirrors human perception and evaluative behavior, we aim to train AI systems that more faithfully reproduce human-like decision-making and perception processes.

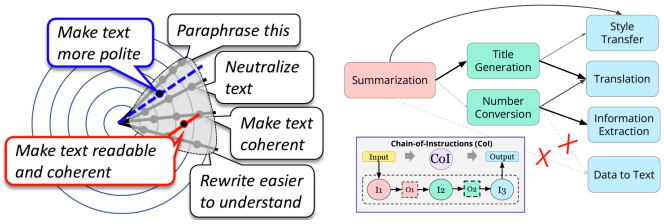
2 Cognitively-aligned AI Models

The goal of *cognitive alignment* is to enhance LLMs' ability to emulate complex human behaviors, like task compositionality, planning, abstraction, reasoning, memory, by either (i) mimicking cognitive processes directly or (ii) incorporating human cognition data into training objectives. In line with Halliday's Systemic Functional Linguistics (SFL) framework [29], our work spans core areas of cognitive capabilities, such as reasoning, planning, and social cognition.



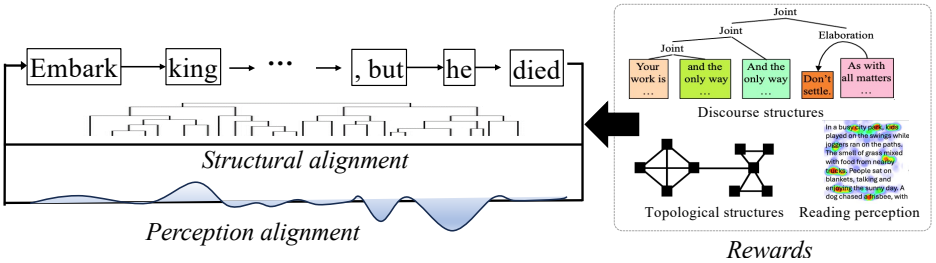
2.1 Reasoning and Task Composition

Human reasoning often bridges incomplete information through external knowledge and iterative refinement. We integrate neural and symbolic systems [40, 17, 39] to fill knowledge gaps while enhancing interpretability [33]. To test and improve compositionality, we created benchmarks for composite and chained tasks, using compositional data augmentation and task-space densification [20, 69, 5]. Human reasoning is often iterative, refining ideas or solutions step-by-step. We mimic this by developing models that emulate iterative reasoning patterns, showing that step-by-step improvements lead to higher-quality, more diverse outputs [16, 47, 15].



2.2 Planning

Humans achieve coherence in writing by ensuring every part of a text fits together to form a complete picture, making structural decisions such as topic choice, sentence order, level of abstraction, and communication strategy. Planning is thus a higher-level cognitive process that involves organizing multiple text passages and hierarchical decision-making process, guiding the surface realization of text based on these plans. We model text planning as hierarchical decision-making, optimizing generation through:

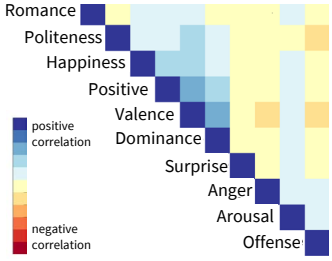


Discourse-Guided Planning. Coherent text is created through planning that aligns sentence sequences with various discourse goals, guided by linguistic theories like rhetorical structure theory (Mann and Thompson, 1988) and script theory (Tomkins, 1978). We develop supervised text planners that incorporate discourse relations [33, 34], topical keywords [36], and social goals inspired by persuasion theories [19]. This approach allows for more controlled, coherent, and interpretable text generation by providing explicit plans (relations, keywords, goals) through human-defined or theory-driven discourse guidance [50, 59, 19].

Self-Supervised Planning and Structural Alignment. Unlike humans, LLM's next-token prediction objective creates a cognitive gap when performing complex writing and planning tasks. We address this by using alignment techniques such as reinforcement learning (RL) to optimize high-level policy decisions

informed by human feedback on generated responses. Our earlier work on self-supervised role-playing framework [31] uses RL to simulate dialogue interactions in recommendation scenarios, and optimize them to achieve the explicit communication goal. Recently, this framework has evolved into "RL with AI Feedback" (RLAIF), where one agent generates responses while another provides explicit feedback, enabling better long-term policy learning between two agents. More recently, we align text structure with human writing via structural rewards [51, 49, 78, 71] and integrate human perception data via dense alignment [55], such as eye-tracking and annotated important lexicons [21, 10, 23, 55] to improve interpretability and generalization.

2.3 Social Cognition



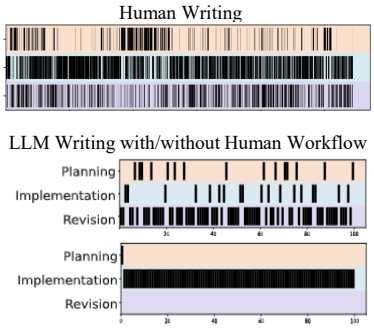
Language encodes interpersonal and societal dimensions of social cognition. Text style emerges from interacting factors such as formality, emotion, and metaphor, reflecting an author's personality and serving specific communicative goals. Understanding these cross-style relationships is key to capturing the nuances of human communication. We built datasets for cross-style analysis, including PASTEL [32] and xSLUE [37], showing that multi-style learning outperforms single-style approaches. Certain style pairs, such as impoliteness–offense, are strongly correlated, while contradictory combinations yield less appropriate outcomes. This underexplored area remains vital for modeling complex stylistic patterns. More recently, we expanded to high-level affective states such as nostalgia [46] and skepticism [64], collecting annotated data for social cognition research. To align models with multiple social aspects, we developed methods for multi-attribute control via data balancing [9], multi-task fine-tuning [37], and policy learning with dynamic reward re-weighting [11].

2.4 Cognitive Efficiency

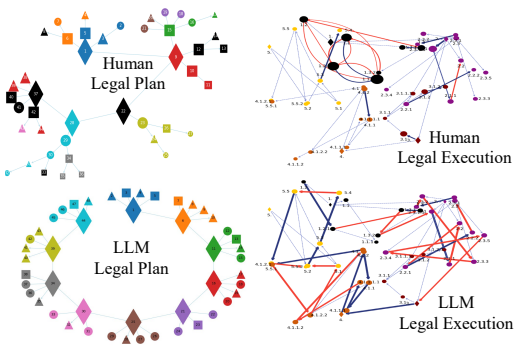
Abstraction enables humans to process knowledge efficiently, bypassing detailed reasoning when shortcuts suffice. Inspired by this, we design algorithms that optimize data usage, computation, and memory in LLM training, reducing hallucinations and improving reliability. Our work includes data-efficient methods that maintain performance with fewer training samples [45, 68], and compute-efficient models which pairs parameters across models via dynamic weight warping [63]. We continue to develop cognitively inspired techniques that optimize LLMs' *cognitive utility function*, balancing performance with resource efficiency [74].

2.5 End-to-end Cognitive Workflow

My long-term goal is to develop cognitively inspired AI systems that can perform complex, multi-step tasks involving causality, abstraction, and memory, while reducing human cognitive load. To build such systems, it is not sufficient to excel in a single cognitive capability. Instead, multiple functions must work in concert, enabling agents to understand the overall workflow of complex and time-consuming tasks (e.g., scientific processes, legal reasoning), and to process high-fidelity tasks quickly and efficiently.



(a) Scholarly Writing Intents over Time

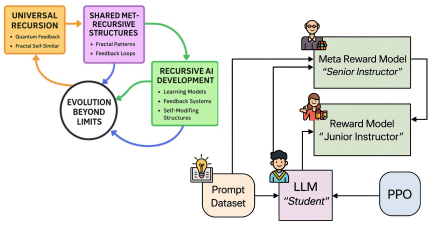


(b) Legal Planning (left) & Execution (right)

Learning from Human-vs-LLM Workflow. The first step is to investigate how human workflows differ from AI workflows and to design methods to align them. Our prior work on collecting long-term human workflow data (in scholarly writing [73] and legal processing [8]) has revealed clear distinctions between

human and AI processes. When end-to-end writing data is used to train models to emulate human writing trajectories, the resulting systems exhibit more human-aligned dynamics and improved writing quality [73, 54]. Similarly, in legal planning and execution, we find that humans often act more spontaneously, sometimes repeating inefficient steps [8], whereas AI tends to rigidly follow pre-defined plans. We observe these contrasting behavioral patterns in human- and AI-based legal workflow data collected in [70], providing key insights for designing cognitively aligned thinking assistants that better bridge workflows between lawyers and AI agents.

Optimizing LLM Workflow On the other hand, we also study how LLMs themselves learn and optimize their cognitive processes during complex tasks. For example, models can refine their own policies or workflows through self-evolving rewards [48] or test-time scaling with verifiable rewards. This line of LLM workflow optimization can ultimately be combined with human workflows to enable more effective collaborative workflow optimization.



Workflow modeling (left) and Meta-Policy Optimization (MPO) [48]

2.6 Extension to Other Modalities

While my work primarily focuses on language, multidisciplinary collaborations have extended it to images and video, showing that multimodal models benefit from complementary cross-modal effects. Effective multimodal learning, however, requires careful design of fusion strategies and cognitive objectives such as reasoning and planning. Key results include:

- Vision–language models improve understanding of global connectivity and graph motif analysis in graph images [6].
- Explicit temporal memory in video LLMs yields more coherent frames [4]; aligning vision–language models with LLM feedback enhances reasoning in video and image tasks [41, 3, 2, 65].
- Integrating multimodal representations into a unified latent space ("Platonic representation") accelerates understanding [25].
- LLMs can plan robotic manipulation tasks from language-based instructions [75].

Beyond multimodality, I also explore (i) autoregressive–diffusion hybrids for text generation [79], (ii) model uncertainty [27, 45, 52] and generalization [1, 28], and (iii) theoretical analyses of learning frameworks [58].

The next research pillars include benchmarking human and machine cognition (§3) and modeling expert workflows to refine AI's cognitive assistance capabilities (§4).

3 Cognitive Benchmarking and Evaluation

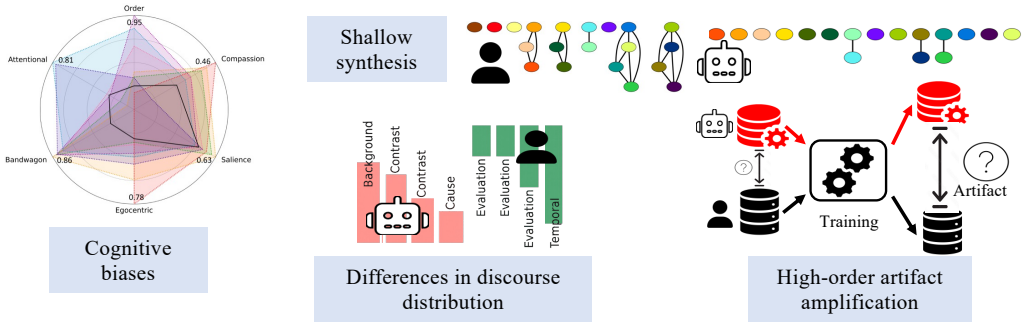
Current LLM evaluations focus on isolated tasks and fail to capture the full cognitive demands of real-world applications. The growing use of LLMs for data collection and evaluation introduces new challenges, including difficulty tracing error sources and amplifying artifacts in AI-generated data. My work addresses these issues through two directions: (i) assessing the potentials and risks of AI-generated data and evaluation (i.e., LLM-as-judge), and (ii) building a comprehensive framework to benchmark human and machine cognition.

3.1 Potentials and Risks of AI-generated Data

Potentials: To reduce annotation costs, we developed data augmentation methods using symbolic rules [17, 40], annotation imputation [61], and information-theoretic measures [44, 45, 68, 52], as well as disagreement-aware sampling [72]. LLMs can generate annotations, prompts, simulated dialogues, and evaluation data, accelerating research and development process. We have leveraged LLM-based evaluations [43, 3] and multi-agent simulations [31, 5, 76] to enrich training data.

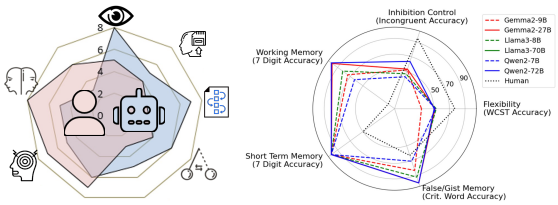
Risks and Biases: Over-reliance on AI-generated data risks creating an "artificial data ecosystem" [7] with:

- *Cognitive Biases* [53] in LLM-based evaluation: e.g., egocentric and length biases [38].
- *Stylistic Discrepancies* [7]: more formal and stylistically distinct from human text.
- *Shallow Synthesis* [62]: shallow citation and knowledge integration.
- *Discourse Biases* [49]: different structural patterns in long-form text.



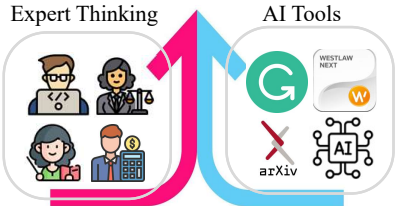
- *Artifact Amplification* [7]: reinforcing biases and artifacts during training.

3.2 Benchmarking of Cognition



Traditional cognitive science experiments provide detailed insights into human cognition but are often small-scale. Aligning LLMs with these fine-grained data [10, 23] has shown promising results, though obtaining high-quality data remains challenging. To scale up these efforts, we are developing a project called CogBench [12], a comprehensive cognitive benchmarking framework. CogBench aims to collect high-quality human cognition data via controlled experiments with cognitive scientists [13], and evaluate human and LLM cognition across selective attention, working memory, reasoning, and reading comprehension [14]. By integrating rigorous data practices with large-scale benchmarking, we aim to establish a more robust, human-centered methodology for assessing and aligning AI capabilities. Our cognitive benchmarking framework grounds AI evaluation in established cognitive science methodologies. This dual focus, applied systems with theoretically grounded benchmarks, ensures advances in both the scientific understanding of human cognition and the design of practical AI tools.

4 Expert-level AI: Assist Expert Thinking Process



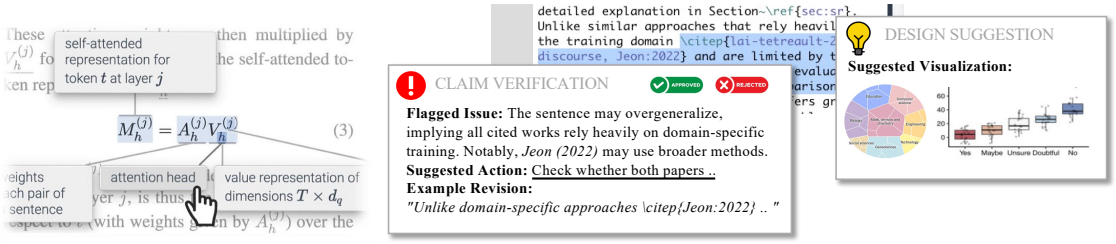
In knowledge-intensive domains such as scientific research and legal practice, a significant cognitive gap remains between human experts and current LLM capabilities. These gaps arise from limitations in handling complex multitasks, incorporating dynamic feedback, and adapting to evolving contexts. Full automation often oversimplifies the nuanced reasoning these tasks require, misaligning AI systems with expert workflows.

My objectives are twofold: (i) to benchmark the skills and knowledge required in expert domains, and (ii) to create collaborative platforms where experts and AI agents interact, exchange feedback, and adapt capabilities over time. This framework aims to benchmark AI cognition in end-to-end professional tasks, identify cognitive gaps, and refine models for closer alignment with human expertise. We first focus on building expert support systems in two domains: *science* and *law*.

4.1 Assistants for Scientists

Reading Assistance In collaboration with AI2 and UC Berkeley, I contributed to the development of *Semantic Reader* [24, 60], which provides in-situ definitions and explanations of terms when reading scientific papers [35, 26], thereby reducing cognitive load. More recently, we launched *SciTalk*, which transforms research papers into short-form videos, making knowledge more accessible to wider audiences [67].

Writing Assistance Scientific writing is a challenging, iterative cognitive task that requires sustained practice, collaboration, and integration of existing literature. I envision the future of scientific writing as a collaborative effort between scientists and AI agents, where both iteratively co-develop high-quality research outputs.



Augmented Reading Interface | Workflow-aware Writing Assistant

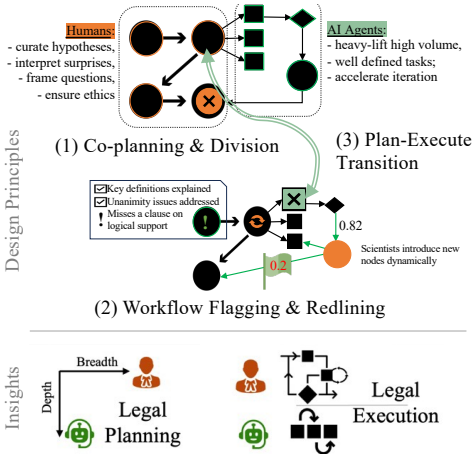
Writing assistance must go beyond supporting the immediate writing context. It should understand the full workflow, capturing the writer's intent and providing in-situ support while minimizing distraction. Also, I focus on building interaction-centric AI systems that evolve through user feedback to better support scientific writing.

Our design principles include:

- *In-Situ Aids for Writing*: Document-level tools that detect overloaded symbols, redundant terminology, and logical inconsistencies using discourse-level modeling for real-time writing support.
- *Workflow-aware Interfaces*: features that flag inconsistent scientific claims or over-toned text, with suggestions. For instance, a scientist drafting a related work section in Overleaf may trigger backend reasoning agents to automatically retrieve cited literature, detect errors or overstatements, and suggest improved citations with inferred rationales.
- *Human-Centric Design and Co-Learning*: Avoiding the "illusion of clarity" (Nguyen, 2021) by carefully designing interfaces that support scientists' thinking processes while reinforcing human–AI collaboration.
- *Learning from Human Workflow*: Collecting human revision and writing workflow data to build cognitively-aligned systems for naturally integrated interactions [15, 73, 57, 76].
- *Learning from Interaction*: Adaptive systems that improve via user feedback, as demonstrated in collaborative editing [15] and iterative taxonomy building [56].
- *Trustworthy Models*: With NSF support, we develop foundation models integrating reliable, cross-modal knowledge for robust writing assistance.

Other Scientific Tasks Beyond reading and writing, scientific workflows involve cognitively demanding stages such as peer review, collaboration, experimentation, and verification. Our lab is developing specialized agents and interfaces to support different stages of scientific tasks. For instance, we created *PeerRead* [30], the first large-scale dataset of papers and reviews from computer science conferences, enabling research on acceptance prediction and aspect-specific review generation. In collaboration with visualization researchers, we are also building systems that automatically generate effective visualizations or insights, adapting interactively to scientists' needs and scaling insights during interaction.

4.2 Assistants for Lawyers



With Open Philanthropy and UMN Law School, we developed *LawFlow* [8], a benchmark for complex legal processing tasks requiring long-horizon reasoning and planning. The dataset captures detailed workflow data from both human lawyers and AI agents, including brainstorming, planning, research, and client communication—for tasks such as "advising a startup." Our findings reveal that human legal reasoning is recursive and exploratory, whereas AI workflows are typically linear and exhaustive. From these human–AI comparisons, we derive design principles for legal interfaces, including collaborative planning, task division, and workflow flagging, and we are actively developing interfaces guided by these insights. Reflections from practicing lawyers [70] further highlight that generative AI already improves junior-lawyer performance, yet adoption lags because evaluating AI outputs is cognitively demanding, especially for less experienced practitioners.

These challenges create several risks: over-trust or over-reliance on AI, or shortcut-driven reasoning (Kosmyna et al., 2025) as well as the potential intensification of expert polarization. Junior lawyers, in particular, risk being sidelined without proper scaffolding. To mitigate these risks, we emphasize the need for tools that up-skill juniors, train them to critically validate AI outputs, and provide workflow-aware assistance that supports lawyers as thinking and co-learning partners, ultimately lowering their cognitive burdens rather than replacing their expertise.

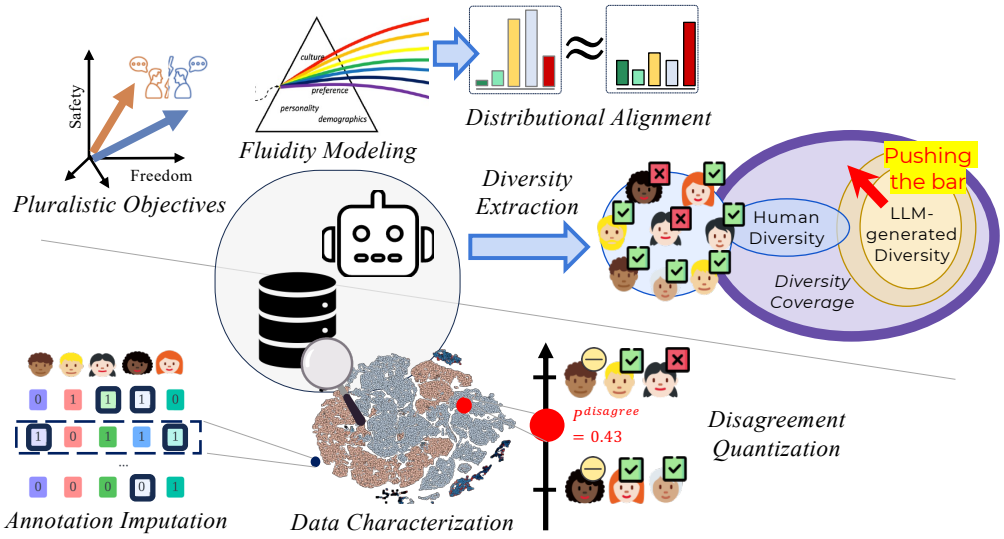
4.3 Computational Advertising and Journalism

As part of my commitment to Naver and MCAL, I contribute to developing computational metrics and algorithms for advertising in both academia and industry. For example, in the advertising market, we model nuanced affective responses such as skepticism [64] and nostalgia [46], moving beyond sentiment-based metrics. Recently, we also study how advertising is shifting from keyword-based search advertising toward generative ad platforms (e.g., Perplexity, Google AI Briefing). We conduct comparative research on search patterns and consumer engagement to better understand the impact of AI on the advertising ecosystem [18]. Finally, I've worked on transforming the workflows of Naver's advertising services by integrating AI technologies. For instance, we are developing ad systems that support offline policy estimation (OPE) in deterministic environment such as ad auctions.

By capturing detailed expert reasoning, we aim to build cognitively inspired models and expert-aware interfaces that adapt to the dynamic cognitive states of professionals. This approach will extend to other domains such as education, medicine, journalism, and programming, forming a broad *Expert Benchmarking* initiative to support specialized AI systems tailored to the cognitive demands of diverse expert roles.

5 Pluralistic Alignment

Our society thrives on diverse perspectives shaped by different backgrounds, identities, and cultures. However, rapid AI advancements often collapse these into a generalized "average," erasing the richness of pluralistic viewpoints. AI must instead promote people-centric technologies that capture and encourage diversity, ensuring inclusivity, ethical growth, and societal alignment. My goal is to build socially aware AI models that represent the full spectrum of human opinions through two complementary approaches: **data-centric** (§5.1) and **model-centric** (§5.2).



5.1 Data-centric Alignment

Defining and capturing diversity is challenging, manifesting at individual, group, and community levels, and varying over time and context. Our "pluralistic representation" approach encodes disagreement and perspective variation directly into datasets. We develop methods to detect, characterize, and augment marginal viewpoints by:

- Predicting missing annotations with collaborative filtering over annotations [61].
- Analyzing annotation properties via learning dynamics [45].

- Quantifying the disagreement level using demographic factors [72].
- Collecting perspectives from language learners to diversify model predictions [77, 42].

5.2 Model-centric Alignment

As NLP tasks grow more subjective, labels often shift from discrete to continuous values. Models that ignore this fluidity risk excluding certain groups. We design pluralistic alignment methods at multiple granularity levels:

- *Individual preferences:* Model disagreement as a proxy for diversity using pairwise preference learning [43].
- *Group distributions:* Extract and align diverse opinions from LLMs [22] with survey-based distributions using distributional alignment.
- *Societal values:* Dynamically combine and resolve value conflicts using crowdsourced resolution scenarios [11].

As AI becomes embedded in education, industry, and politics, it must reflect diversity, plurality, and mutual respect. Without this, systems risk amplifying monolithic and biased perspectives, increasing polarization. Pluralistic AI seeks to model coexistence among differing values, beliefs, and cultural backgrounds, fostering inclusive and socially responsible AI that values and encourages diverse perspectives. My future work will operationalize pluralistic alignment through quantitative diversity metrics, disagreement-aware training objectives, and cross-cultural evaluation datasets.

References

[1] Dyah Adila and Dongyeop Kang. Understanding out-of-distribution: A perspective of data dynamics. In *ICBINB@NeurIPS*, 2021.

[2] Daechul Ahn, Yura Choi, San Kim, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. Isr-dpo: Aligning large multimodal models for videos by iterative self-retrospective dpo. In *Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[3] Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. Tuning large multimodal models for videos using reinforcement learning from ai feedback. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[4] Daechul Ahn, Daneul Kim, Gwangmo Song, Seung Hwan Kim, Honglak Lee, Dongyeop Kang, and Jonghyun Choi. Story visualization by online text augmentation with context memory. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2023.

[5] Hyungjoo Chae, Yongho Song, Kai Tzu iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[6] Debarati Das, Ishaan Gupta, Jaideep Srivastava, and Dongyeop Kang. Which modality should I use - text, motif, or image? : Understanding graphs with large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 503–519, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[7] Debarati Das, Karin De Langis, Anna Martin-Boyle, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Anugrah Hayati, Risako Owan, Bin Hu, Ritik Parkar, Ryan Koo, Jonginn Park, Aahan Tyagi, Libby Ferland, Sanjali Roy, Vincent Liu, and Dongyeop Kang. Under the surface: Tracking the artifactuality of llm-generated data, 2024.

[8] Debarati Das, Khanh Chi Le, Ritik Sachin Parkar, Karin De Langis, Brendan Madson, Chad M. Berryman, Robin M Willis, Daniel H Moses, Brett McDonnell, Daniel Schwarcz, and Dongyeop Kang. Lawflow : Collecting and simulating lawyers' thought processes. In *Conference on Language Modeling (COLM)*, 2025.

[9] Debarati Das, David Ma, and Dongyeop Kang. Balancing effect of training dataset distribution of multiple styles for multi-style text transfer. In *Findings of the Annual Meeting of the Association for Computational Linguistics (ACL) Findings*, 2023.

[10] Karin de Langis and Dongyeop Kang. A comparative study on textual saliency of styles from eye tracking, annotations, and language models. In Jing Jiang, David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 108–121, Singapore, December 2023. Association for Computational Linguistics.

[11] Karin de Langis, Ryan Koo, and Dongyeop Kang. Dynamic multi-reward weighting for multi-style controllable generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[12] Karin de Langis, Jong Inn Park, Bin Hu, Khanh Chi Le, Andreas Schramm, Michael C. Mensink, Andrew Elfenbein, and Dongyeop Kang. A framework for robust cognitive evaluation of llms, 2025. under review.

[13] Karin de Langis, Jong Inn Park, Andreas Schramm, Bin Hu, Khanh Chi Le, and Dongyeop Kang. How llms comprehend temporal structure in narratives: A case study in cognitive evaluation of llms. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[14] Karin de Langis, William Walker, Rimika Dhara, and Dongyeop Kang. Understanding reading behaviors during the preference annotation task, 2025. under review.

[15] Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 96–108, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[16] Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590, Dublin, Ireland, May 2022. Association for Computational Linguistics (ACL).

[17] Steven Y. Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. Genaug: Data augmentation for finetuning text generators. In *Deep Learning Inside Out (DeeLIO) Workshop at EMNLP 2020*, Online, November 2020.

[18] Gabriel Garlough-Shah, Jong Inn Park, Shirley Anugrah Hayati, Dongyeop Kang, and Jisu Huh. Consumer engagement with ai-powered search engines and implications for the future of search advertising. In *Association for Education in Journalism and Mass Communication (AEJMC)*, 2024.

[19] Shirley A. Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. Inspired: Toward sociable recommendation dialog systems. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November 2020.

[20] Shirley Anugrah Hayati, Taehee Jung, Tristan Bodding-Long, Sudipta Kar, Abhinav Sethy, Joo-Kyung Kim, and Dongyeop Kang. Chain-of-instructions: Compositional instruction tuning on large language models, 2024.

[21] Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. Does bert learn as humans perceive? understanding linguistic styles through lexica. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[22] Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. How far can we extract diverse perspectives from large language models? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[23] Shirley Anugrah Hayati, Kyumin Park, Dheeraj Rajagopal, Lyle Ungar, and Dongyeop Kang. Stylex: Explaining styles with lexicon-based human perception. In *European Chapter of Association for Computational Linguistics (EACL)*, 2023.

[24] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2021.

[25] Minoh Jeong, Min Namgung, Zae Myung Kim, Dongyeop Kang, Yao-Yi Chiang, and Alfred Hero. Anchors aweigh! sail for optimal unified multi-modal representations, 2025. under review.

[26] Hwiyeol Jo, Dongyeop Kang, Andrew Head, and Marti A. Hearst. Modeling mathematical notation semantics in academic papers. In *Findings on Empirical Methods in Natural Language Processing (EMNLP Findings)*, 2021.

[27] Taehee Jung, Dongyeop Kang, Hua Cheng, Lucas Mentch, and Thomas Schaaf. Posterior calibrated training on sentence classification tasks. In *2020 Annual Conference of the Association for Computational Linguistics (ACL)*, 2020.

[28] Taehee Jung, Joo-Kyung Kim, Sungjin Lee, and Dongyeop Kang. Cluster-guided label generation in extreme multi-label classification. In *European Chapter of Association for Computational Linguistics (EACL)*, 2023.

[29] Dongyeop Kang. *Linguistically Informed Language Generation: A Multifaceted Approach*. PhD dissertation, Carnegie Mellon University, Pittsburgh, PA, May 2020. Available at https://kilthub.cmu.edu/articles/thesis/Linguistically_Informed_Language_Generation_A_Multi-faceted_Approach/24990648.

[30] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, New Orleans, USA, June 2018.

[31] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, November 2019.

[32] Dongyeop Kang, Varun Gangal, and Eduard Hovy. (male, bachelor) and (female, ph.d) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, November 2019.

[33] Dongyeop Kang, Varun Gangal, Ang Lu, Zheng Chen, and Eduard Hovy. Detecting and explaining causes from text for a time series event. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2017.

[34] Dongyeop Kang, Hiroaki Hayashi, Alan W Black, and Eduard Hovy. Linguistic versus latent relations for modeling coherent flow in paragraphs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, November 2019.

[35] Dongyeop Kang, Andrew Head, Risham Sidhu, Kyle Lo, Daniel Weld, and Marti A. Hearst. Document-level definition detection in scholarly documents: Existing models, error analyses, and future directions. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 196–206, Online, November 2020. Association for Computational Linguistics.

[36] Dongyeop Kang and Eduard Hovy. Plan ahead: Self-supervised text planning for paragraph completion. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November 2020.

[37] Dongyeop Kang and Eduard Hovy. Style is NOT a single variable: Case studies for cross-stylistic language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2376–2387, Online, August 2021. Association for Computational Linguistics (ACL).

[38] Dongyeop Kang*, Taehee Jung*, Lucas Mentch, and Eduard Hovy. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, November 2019.

[39] Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Peter Clark. Bridging knowledge gaps in neural entailment via symbolic models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, November 2018.

[40] Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard Hovy. Adventure: Adversarial training for textual entailment with knowledge-guided examples. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, July 2018.

[41] Jihyung Kil, Farideh Tavazoee, Dongyeop Kang, and Joo-Kyung Kim. Ii-mmr: Identifying and improving multimodal multi-hop reasoning in visual question answering. In *Annual Meeting of the Association for Computational Linguistics (ACL) Findings*, 2024.

[42] Haechan Kim, Junho Myung, Seoyoung Kim, Sungpah Lee, Dongyeop Kang, and Juho Kim. Learnervoice: A dataset of non-native english learners' spontaneous speech. In *Conference of the International Speech Communication Association (Interspeech)*, 2024.

[43] Jaehyung Kim, Shin Jinwoo, and Dongyeop Kang. Prefer to classify: Improving text classifiers via auxiliary preference learning. In *International Conference on Machine Learning (ICML)*, 2023.

[44] Jaehyung Kim, Dongyeop Kang, Sungsoo Ahn, and Jinwoo Shin. What makes better augmentation strategies? augment difficult but not too different. In *International Conference on Learning Representations (ICLR)*, 2022.

[45] Jaehyung Kim, Yekyung Kim, Karin de Langis, and Dongyeop Kang. infoverse: A universal framework for dataset characterization with multidimensional meta-information. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[46] Katie Kim, Bugil Chang, Debarati Das, Smitha Sudheendra, Jisu Huh, Dongyeop Kang, and Jaideep Srivastava. Rebuilding social connection and enhancing advertising effects through the nostalgic appeal during the pandemic. In *Association for Education in Journalism and Mass Communication (AEJMC)*, 2023.

[47] Zae Myung Kim, Wanyu Du, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. Improving iterative text revision by learning where to edit from other revision tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9986–9999, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[48] Zae Myung Kim, Chanoo Kim, Vipul Raheja, and Dongyeop Kang. Toward evaluative thinking: Meta policy optimization with evolving reward models, 2025. under review.

[49] Zae Myung Kim, Kwang Hee Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. Threads of subtlety: Detecting machine-generated texts through discourse motifs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[50] Zae Myung Kim, Vassilina Nikoulina, Dongyeop Kang, Didier Schwab, and Laurent Besacier. Visualizing Cross-Lingual discourse relations in multilingual TED corpora. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 165–170, Punta Cana, Dominican Republic and Online, November 2021. Association for Computational Linguistics.

[51] Zae Myung Kim, Anand Ramachandran, Farideh Tavazoee, Joo-Kyung Kim, Oleg Rokhlenko, and Dongyeop Kang. Align to structure: Aligning large language models with structural information, 2025. under review.

[52] Ryan Koo, Yekyung Kim, Dongyeop Kang, and Jaehyung Kim. Meta-crafting: Improved detection of out-of-distributed texts via crafting metadata space (student abstract). In *Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI) Student Abstract*, 2024.

[53] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Annual Meeting of the Association for Computational Linguistics (ACL) Findings*, 2024.

[54] Ryan Koo, Anna Martin, Linghe Wang, and Dongyeop Kang. Decoding the end-to-end writing trajectory in scholarly manuscripts. In *Proceedings of the Second Workshop on Intelligent and Interactive Writing Assistants (In2Writing) at CHI 2023*, 2023. https://minnesotanlp.github.io/REWARD_demo/.

[55] Ryan Koo, Ian Yang, Vipul Raheja, Mingyi Hong, Kwang-Sung Jun, and Dongyeop Kang. Learning explainable dense reward shapes via bayesian optimization, 2025. under review.

[56] Minhwa Lee, Zae Myung Kim, Vivek Khetan, Alex Kass, and Dongyeop Kang. Cotaxon: Enhancing domain expertise through human-ai collaborative taxonomy construction, 2025. under review.

[57] Minhwa Lee, Zae Myung Kim, Vivek A. Khetan, and Dongyeop Kang. Human-ai collaborative taxonomy construction: A case study in profession-specific writing assistants. In *CHI 2024 In2Writing Workshop*, 2024.

[58] Chenliang Li, Siliang Zeng, Zeyi Liao, Jiaxiang Li, Dongyeop Kang, Alfredo Garcia, and Mingyi Hong. Joint reward and policy learning with demonstrations and human feedback improves alignment. In *International Conference on Learning Representations (ICLR)*, 2025.

[59] Chu-Cheng Lin, Dongyeop Kang, Michael Gamon, Madian Khabsa, Ahmed Hassan Awadallah, and Patrick Pantel. Actionable email intent modeling with reparametrized rnn. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[60] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu Kang, Egor Klevak, Bailey Kuehl, Michael Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. The semantic reader project: Augmenting scholarly documents through ai-powered interactive reading interfaces. In *Communications of ACM*, 2023.

[61] London Lowmanstone, Ruyuan Wan, Risako Owan, Jaehyung Kim, and Dongyeop Kang. Annotation imputation to individualize predictions: Initial studies on distribution dynamics and model predictions. In *Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ECAI 2023*, 2023.

[62] Anna Martin-Boyle, Aahan Tyagi, Marti A. Hearst, and Dongyeop Kang. Shallow synthesis of knowledge in gpt-generated texts: A case study in automatic related work composition, 2024.

[63] James Mooney, Vipul Raheja, Vivek Kulkarni, Ryan Koo, and Dongyeop Kang. Pairlens: Detecting correspondences in paired models, 2025. under review.

[64] Smitha Muthya Sudheendra, Maral Abdollahi, Dongyeop Kang, Jisu Huh, and Jaideep Srivastava. SkOTaPA: A dataset for skepticism detection in online text after persuasion attempt. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14871–14876, Torino, Italia, May 2024. ELRA and ICCL.

[65] Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. Zero-shot natural language video localization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021.

[66] Rose Neis, Karin de Langis, Zae Myung Kim, and Dongyeop Kang. An analysis of reader engagement in literary fiction through eye tracking and linguistic features. In *Workshop on Narrative Understanding (WNU) @ACL 2023*, 2023.

[67] Jong Inn Park, Maanas Taneja, Qianwen Wang, and Dongyeop Kang. Stealing creator's workflow: A creator-inspired agentic framework with iterative feedback loop for improved scientific short-form generation, 2025. under review.

[68] Ritik Sachin Parkar, Jaehyung Kim, Jong Inn Park, and Dongyeop Kang. Selectllm: Can llms select important instructions to annotate?, 2024.

[69] Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. Coedit: State-of-the-art text editing by task-specific instruction tuning. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP Findings)*, 2023.

[70] Daniel Schwarcz, Debarati Das, Dongyeop Kang, and Brett McDonnell. Thinking like a lawyer in the age of generative ai: Cognitive limits on ai adoption among lawyers, 2025. under review.

[71] Zhecheng Sheng, Tianhao Zhang, Chen Jiang, and Dongyeop Kang. Bbscore: A brownian bridge based metric for assessing text coherence. In *Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[72] Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. Everyone's voice matters: Quantifying and distinguishing subjective annotation disagreement using demographic information. In *Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[73] Linghe Wang, Minhwa Lee, Ross Volkov, Luan Chau, and Dongyeop Kang. Scholawrite: A dataset of end-to-end scholarly writing process, 2025. under review.

[74] Quan Wei, Chung-Yiu Yau, Hoi To Wai, Yang Zhao, Dongyeop Kang, Youngsuk Park, and Mingyi Hong. Roste: An efficient quantization-aware supervised fine-tuning approach for large language models. In *International Conference on Machine Learning (ICML)*, 2025.

[75] Carl Winge, Adam Imdieke, Bahaa Aldeeb, Dongyeop Kang, and Karthik Desingh. Talk through it: End user directed manipulation learning. In *IEEE Robotics and Automation Letters (RA-L)*, 2024.

[76] Ruixin Yang, Dheeraj Rajagopal, Shirley Anugrah Hayati, Bin Hu, and Dongyeop Kang. Confidence calibration and rationalization for LLMs via multi-agent deliberation. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024.

[77] Haneul Yoo, Rifki Afina Putri, Changyoon Lee, Youngin Lee, So-Yeon Ahn, Dongyeop Kang, and Alice Oh. Rethinking annotation: Can language learners contribute?, 2023.

[78] Tianhao Zhang, Zhexiao Lin, Zhecheng Sheng, Chen Jiang, and Dongyeop Kang. On the sequence evaluation based on stochastic processes, 2024.

[79] Hao Zou, Zae Myung Kim, and Dongyeop Kang. A survey of diffusion models in natural language processing, 2023.