In today's rapidly evolving AI landscape, machine learning models, including foundation models, increasingly shape how people access information, form judgments, communicate, learn, and engage with digital environments. Yet despite impressive capabilities, these systems frequently fail in ways that reveal a gap between ***model capabilities*** and ***human trust***. For instance, AI systems may be presented as reliable knowledge sources yet produce confidently incorrect or fabricated outputs, including hallucinated facts that compromise decision making [1]. Trust also falters when chatbots ignore social norms, raising public concern about unreliable or inappropriate model responses, especially for young teenagers [2]. Clinicians similarly hesitate to adopt AI diagnostic tools because they lack transparency, have little clinical validation, and may perform inconsistently across patient groups [3]. Addressing these issues is not simply a technical necessity, but a foundational requirement for developing AI systems that remain reliably aligned with both human needs and societal values.

My research vision, as a crucial part of *trustworthy & human-centered AI*, is to **develop the scientific foundations of human-AI alignment so that AI systems can understand and act upon human intent, information behavior, and social context.** I formalize this perspective by viewing alignment as jointly minimizing *task-related error* and *human-centered misalignment* through a unified objective:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{task}} + \lambda \, \mathcal{L}_{\text{human}}, \tag{1}$$

where $\mathcal{L}_{\text{task}}$ captures utility or task performance, and $\mathcal{L}_{\text{human}}$ represents misalignment with human intent and social values. Crucially, $\mathcal{L}_{\text{human}}$ should be instantiated at multiple scales, including the *individual* (e.g., addressing user intent multi-modality and ambiguity), *group* (e.g., alleviating representation and exposure bias), and *societal* levels (e.g., supporting social norms), providing a coherent lens for interpreting alignment across human contexts.

However, effectively instantiating $\mathcal{L}_{\text{human}}$ at scale is challenging because human feedback—whether expressed through clicks, edits, demonstrations, or deliberation signals—is heterogeneous, indirect, and often noisy. Such signals rarely provide explicit labels for alignment, making it insufficient to treat them as simple supervision. To meet this challenge, my research develops foundational algorithms that model human feedback as a structured signal rather than a set of isolated labels, enabling AI systems to uncover the latent objectives that guide human behavior. Through this lens, I design models that remain *reliable* under ambiguity, *robust* to noise and bias, and *aligned* with both individual and societal needs. I operationalize this vision through two complementary pillars spanning **micro- and macro-level human-AI alignment**, as illustrated in Figure 1.

- **Human-AI Alignment for Individual Needs (Micro-level)** — I develop methods to model evolving user preferences expressed through implicit or ambiguous *information behavior* and interactions with digital content. This includes techniques for understanding ambiguous and multi-faceted preferences through contrastive and multi-feedback modeling [4, 5, 6, 7], learning from imperfect and noisy feedback [8, 9], and utilizing external knowledge and sequential information signals to enhance personalization under weak supervision [10, 11]. These methods enable AI systems to interpret user intent accurately while adapting to evolving information needs and contextual nuances.
- **Human-AI Alignment for Societal Impact (Macro-level)** — I extend alignment to high-stakes and multi-actor settings such as education, public platforms, and collaborative decision-making. My work introduces fairness-aware optimization to balance multi-stakeholder objectives [12, 13, 14], transparent representations that enable user-directed *knowledge organization* and scrutiny [15], and AI tutors that respond to both cognitive and emotional information cues [16, 17, 18]. These efforts help shape responsible *information ecosystems* in which AI behaves not only effectively, but also ethically and contextually within real-world social systems.

My research has led to impactful publications in top-tier venues such as NeurIPS, ICLR, ICML, The Web Conference, SIGIR, EMNLP, TMLR, TKDE, CHI, and CSCW, and has been recognized with honors including the **McGill Graduate Excellence Award**, the **Borealis AI Fellowship** *(10 students across Canada)*, and the **Fonds de recherche du Québec Doctoral Fellowship** *(ranked 1st in Quebec)*. My research experience also includes internships at Google Research, Google DeepMind, and Microsoft Research, where I addressed *human-centered, real-world challenges* in natural language processing, personalization, and knowledge modeling. My first-authored research on multi-interest user modeling (NeurIPS 2024) [4] has been deployed in Google YouTube Music's recommendation system, and my work on reliable structured information extraction (EMNLP 2024, Oral) [19] has

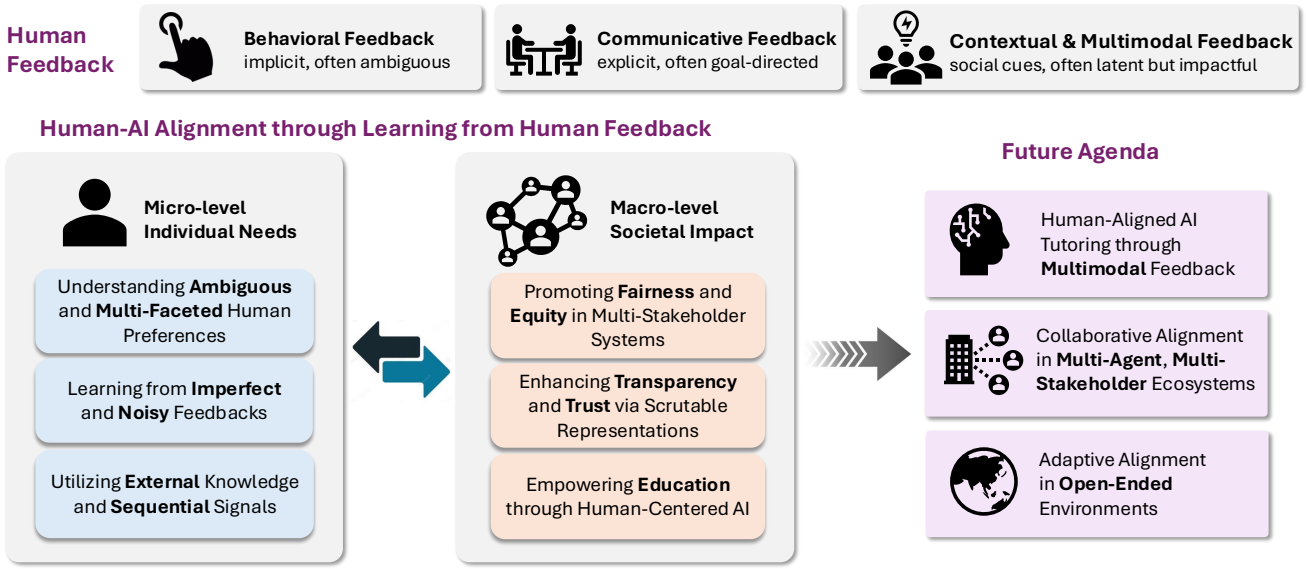

Figure 1: Overview of my research on Human-AI Alignment through learning from human feedback.

been integrated into the Alexandria knowledge base construction pipeline at Microsoft Research. These experiences have shaped my commitment and continue to motivate my pursuit of ***human-centric, socially grounded AI*** in my academic career.

Human-AI Alignment for Individual Needs

Motivation. As AI systems are increasingly deployed to support human goals across domains such as communication, education, search, and decision-making, it becomes essential to model not only what users do but what they intend. Yet human behavior and information work are often ambiguous, context-dependent, and conveyed through implicit, partial, or noisy signals. These challenges are amplified in settings involving diverse individuals or groups, where information needs and preferences may shift over time or conflict with one another. Without principled models for reasoning under such ambiguity and uncertainty, AI systems risk misinterpreting user intent, overfitting to surface-level behaviors, and failing to adapt as preferences evolve. To address this, my research develops foundational methods for modeling multi-faceted user intent, learning from imperfect feedback, and incorporating external signals, ultimately enabling more robust, adaptive, and human-aligned AI systems.

Understanding Ambiguous and Multi-Faceted Human Preferences. Human preferences are often ambiguous, context-dependent, and only partially observable. My research develops methods that disentangle and adapt to such complexity by modeling intent as multi-faceted and uncertain. In *IMCAT* [5] (ICDE 2023), I proposed a contrastive learning framework that aligns user interactions with tag-level metadata to derive *intent-aware embeddings*. This improves both interpretability and prediction. A follow-up work [6] (TOIS 2023) captures dependencies across different types of implicit feedback via coarse-to-fine matching, enabling richer multi-interest modeling. To move beyond fixed-vector user embeddings, I introduced *density-based user representation* [4] (NeurIPS 2024), modeling users as distributions over intents using Gaussian Process Regression. This captures preference diversity and supports robust retrieval under sparse or heterogeneous feedback. At the group level, I further proposed an *ambiguity-sensitive metric* [7] (SIGIR 2024) that accounts for disagreement among users, revealing the need for fundamentally different modeling strategies beyond individual-level aggregation. *These studies treat human intent not as a single point but as a structured, variable signal shaped by context and interaction patterns.*

Learning from Imperfect and Noisy Feedback. In many real-world scenarios, user feedback is implicit, partial, or noisy—posing challenges for accurately inferring intent. My research develops adaptive learning frameworks that treat such signals not as noise to suppress, but as informative cues to be modeled and leveraged. My work [8] (CIKM 2022) introduces a bilevel optimization framework that *reweights training samples* based on contextual informativeness, selectively emphasizing more meaningful comparisons to improve generalization. Most recently, we proposed *Plugin* [9] (ICML 2025), framing LLM adaptation as a *label noise correction* problem. By reweighting token-level output logits using lightweight, task-specific signals, *Plugin* supports effective personalization to downstream tasks without gradient access or model fine-tuning. *These efforts show that imperfect feedback carries*

structured information about relevance and reliability, enabling more robust alignment under real-world conditions.

Utilizing External Knowledge and Sequential Signals. To better align AI systems with nuanced human preferences under weak supervision, I explore methods that incorporate external knowledge and model sequential user behavior. My work *Hyper-Know* [10] (AAAI 2021) embeds item-side knowledge graphs in *hyperbolic space*, preserving hierarchical relations while enabling fine-grained attention over semantically related entities. This enhances recommendation quality, especially in sparse or long-tail scenarios where user feedback alone may be insufficient. Building upon this, my work [11] (CIKM 2024) proposes a context-aware contrastive learning framework for sequential behavior modeling. It uses a diffusion-based sampling strategy to generate *semantically consistent augmented sequences*, aligning user representations with evolving intent patterns and capturing subtle shifts in preference over time. *These approaches reveal that external knowledge and temporal context are essential signals for interpreting user intent beyond what direct feedback can provide.*

Human-AI Alignment for Societal Impact

Motivation. As AI systems mediate decisions that affect not only individuals but also communities, institutions, and broader information ecosystems, it becomes essential to consider the societal consequences of system behavior. Real-world deployments involve multiple stakeholders—users, content providers, platforms—whose goals may diverge or conflict. In such settings, alignment must extend beyond optimizing for individual preferences to incorporate social values such as fairness, inclusivity, and transparency. High-impact domains like education further illustrate how cognitive, emotional, and social processes shape system effectiveness and public trust. These challenges reveal that alignment is inherently a multi-actor and multi-level problem, requiring models that account for interactions among people, institutions, and AI components. My research expands human-AI alignment from individual preference modeling to the societal dimensions of AI design, contributing to systems that are *equitable*, *transparent*, and *trustworthy* within complex information environments.

Promoting Fairness and Equity in Multi-Stakeholder Systems. My research embeds fairness as a core design principle in multi-stakeholder AI systems. My work, *JMEFairness* [12] (SIGIR 2022), formalizes a framework for exposure fairness that considers the perspectives of both users and content providers. Rather than treating exposure disparities in isolation, this framework introduces multi-axis metrics that account for systemic imbalances—such as representational or allocative harm—arising from asymmetric visibility across demographic groups. To operationalize fairness in real-world multi-stakeholder environments, I developed a multi-objective optimization framework [13] (TOIS 2022) that reconciles competing goals across stakeholders—such as user satisfaction, creator visibility, and platform utility—under a unified learning objective. Leveraging smooth ranking techniques and differentiable fairness constraints, the framework balances accuracy with stakeholder equity through *Pareto-efficient trade-offs*, enabling dynamic policy selection based on collective welfare considerations. *These efforts demonstrate that fairness in information ecosystems must be treated as a system-level property shaped by interactions among multiple actors, rather than as a correction applied to isolated components.*

Enhancing Transparency and Trust via Scrutable Representations. Trust in AI systems is closely tied to their ability to communicate reasoning in a form that users can understand and act upon. To this end, we developed a framework, *TEARS* [15] (WWW 2025), for generating *transparent user representations*—natural language summaries that reflect user preferences and can be edited to guide model output. Using recommender systems as a testbed, we align interpretable summaries with latent collaborative filtering signals via optimal transport, enabling high-quality predictions that remain transparent and controllable. This hybrid design allows users to steer recommendations through simple edits, fostering a more trustworthy and transparent interaction between users and AI systems. *These results show that transparency can be achieved not by simplifying models, but by structuring representations so that human understanding and model reasoning operate in the same controllable space.*

Empowering Education through Human-Centered AI. Education presents a fertile context for aligning AI with human values, particularly through supportive, reflective, and socially grounded interactions. Our work explores how language models can scaffold learning across individual and collaborative settings. We developed *SSRLBot* [17] (AIED 2025), an LLM-based agent that leverages socially shared regulation theory to analyze diagnostic team interactions and provide contextualized feedback for medical education. Complementing this system, our multimodal study on medical simulation [20] (CSCW 2025) analyzes how facial-expression-derived emotions

co-occur with socially shared regulation behaviors, revealing distinct affective–cognitive patterns across novice and expert learners and highlighting the need for adaptive, emotionally aware scaffolding. In parallel, we analyzed the co-occurrence of physiological signals and learners' diagnostic dialogues [18] (AIED 2025) to identify *critical moments in team decision-making*, offering implications for how educators and AI tutors can better respond to synchronous interactions. Complementing these systems, our study of social annotation in collaborative reading [16] (CHI 2024) reveals how peer acknowledgments foster engagement and psychological safety. *These efforts contribute to a broader vision of AI's educational impact that adapts to learner needs while promoting social support, reflective practice, and pedagogical alignment.*

Future Agenda

As AI systems become deeply embedded in education, healthcare, scientific discovery, and public digital ecosystems, ensuring that they remain aligned with human intent and social values requires more than static optimization. Future AI systems must learn from multimodal human feedback, coordinate across multiple agents and stakeholders, and adapt to evolving norms over time. Building on my work at the intersection of human-centered modeling, preference learning, and multi-agent system alignment, I aim to develop the foundations of *trustworthy and scalable* human-AI alignment across the following directions. Each direction can be viewed as instantiating $\mathcal{L}_{\text{human}}$ in a new structural form, whether through multimodal signals, multi-agent feedback, or temporal value dynamics, providing a unified framework for advancing the science of human-AI alignment.

Human-Aligned AI Tutoring through Multimodal Feedback. AI has been increasingly used to support learning and instruction, making education an important area for studying human-AI alignment. Educational environments provide rich signals about human cognition, emotion, and intention [21, 22]. Building on my work in instructional modeling and collaborative learning systems [16, 20, 18, 17], I aim to develop *AI tutors that learn from multimodal feedback*, including textual edits, voice cues, gaze, gestures, and interaction traces. I will combine edit-based learning with interpretable student models that track learning state, uncertainty, and engagement, and explore fusion architectures that integrate vision, language, and paralinguistic cues to infer pedagogical signals such as confusion or frustration. More broadly, this direction advances the foundations of multimodal human feedback modeling, essential for alignment in domains involving rich, ambiguous human signals.

Collaborative Alignment in Multi-Agent, Multi-Stakeholder Ecosystems. Real-world AI systems act within ecosystems of multiple humans and multiple agents with distinct incentives and value systems. Building on my research in group-aware evaluation and fairness-aware optimization [13, 12, 7], I aim to develop *collaborative alignment protocols* that operate across human–human, human–AI, and AI–AI interactions. This includes (i) formalizing role-sensitive feedback provenance, (ii) designing deliberative workflows—protest edits, voting, debate-based aggregation—to enable structured oversight, and (iii) developing negotiation and delegation mechanisms grounded in game theory and influence diagrams. Extending system-level optimization frameworks such as *Sys-DPO* [23], I will explore coordinated alignment across interacting components. These efforts reflect the core insight that alignment must be understood not only at the individual level but also as an emergent property of interactions among multiple agents and stakeholders.

Adaptive Alignment in Open-Ended and Evolving Environments. In high-stakes domains such as scientific inquiry, policy, and law, ground truth is fluid and shaped by evolving societal norms and expert practices [24, 25]. Building on my work in feedback-aware modeling and model-based optimization [8, 4, 26], I aim to develop *adaptive alignment frameworks* that allow AI systems to remain aligned as preferences and values shift over time. This includes learning preference trajectories via latent state-space models, integrating human edits into continuous adaptation loops, and incorporating deliberative signals such as upvotes, critiques, or arguments into long-horizon alignment objectives. I will also investigate stability–plasticity trade-offs to ensure systems retain past value structures while absorbing new norms. This direction treats alignment as a dynamic objective—necessary for AI systems operating responsibly in open-ended human environments.

Together, these directions aim to establish a principled framework for human-AI alignment that is multimodal, collaborative, and adaptive. By integrating disciplinary insights with scalable learning algorithms and multi-agent reasoning, my long-term vision is to build AI systems that participate responsibly in complex human environments, supporting individual growth, cooperative decision-making, and socially grounded intelligence.

References

[1] Yujie Sun, Dongfang Sheng, Zihan Zhou, and Yifei Wu. "AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content". In: *Humanities and Social Sciences Communications* 11.1 (2024), pp. 1–14.

[2] Laura Kuenssberg. "'A predator in your home': Mothers say chatbots encouraged their sons to kill themselves". In: *BBC News* (Nov. 2025). URL: https://www.bbc.com/news/articles/ce3xgwyywe4o.

[3] Saleh Afroogh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. "Trust in AI: progress, challenges, and future directions". In: *Humanities and Social Sciences Communications* 11.1 (2024), pp. 1–30.

[4] **Haolun Wu**, Ofer Meshi, Masrour Zoghi, Fernando Diaz, Xue Steve Liu, Craig Boutilier, and Maryam Karimzadehgan. "Density-based User Representation using Gaussian Process Regression for Multi-interest Personalized Retrieval". In: *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*. Vol. 37. 2024, pp. 52568–52594.

[5] **Haolun Wu**, Yingxue Zhang, Chen Ma, Wei Guo, Ruiming Tang, Xue Liu, and Mark Coates. "Intent-aware multi-source contrastive alignment for tag-enhanced recommendation". In: *IEEE 39th international conference on data engineering (ICDE)*. IEEE. 2023, pp. 1112–1125.

[6] Chang Meng, Ziqi Zhao, Wei Guo, Yingxue Zhang, **Haolun Wu**, Chen Gao, Dong Li, Xiu Li, and Ruiming Tang. "Coarse-to-fine knowledge-enhanced multi-interest learning framework for multi-behavior recommendation". In: *ACM Transactions on Information Systems (TOIS)*. 42.1 (2023), pp. 1–27.

[7] **Haolun Wu**, Bhaskar Mitra, and Nick Craswell. "Towards group-aware search success". In: *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 2024, pp. 123–131.

[8] **Haolun Wu**, Chen Ma, Yingxue Zhang, Xue Liu, Ruiming Tang, and Mark Coates. "Adapting triplet importance of implicit feedback for personalized recommendation". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*. ACM, 2022, pp. 2148–2157.

[9] Gaurush Hiranandani*, **Haolun Wu***, Subhojyoti Mukherjee, and Sanmi Koyejo. "Logits are All We Need to Adapt Closed Models". In: *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. * correspondence. 2025.

[10] Chen Ma, Liheng Ma, Yingxue Zhang, **Haolun Wu**, Xue Liu, and Mark Coates. "Knowledge-enhanced top-k recommendation in poincaré ball". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 5. 2021, pp. 4285–4293.

[11] Ziqiang Cui, **Haolun Wu**, Bowei He, Ji Cheng, and Chen Ma. "Diffusion-based Contrastive Learning for Sequential Recommendation". In: *Proceedings of the 33rd ACM International Conference on Information & Knowledge Management (CIKM)*. ACM, 2024, pp. 404–414.

[12] **Haolun Wu**, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. "Joint multisided exposure fairness for recommendation". In: *Proceedings of the 45th International ACM SIGIR Conference on research and development in information retrieval*. 2022, pp. 703–714.

[13] **Haolun Wu**, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. "A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation". In: *ACM Transactions on Information Systems (TOIS)*. 41.2 (2022), pp. 1–29.

[14] **Haolun Wu**, Yansen Zhang, Chen Ma, Fuyuan Lyu, Bowei He, Bhaskar Mitra, and Xue Liu. "Result diversification in search and recommendation: A survey". In: *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. IEEE, 2024.

[15] Emiliano Penaloza, Olivier Gouvert, **Haolun Wu**, and Laurent Charlin. "TEARS: Text Representations for Scrutable Recommendations". In: *Proceedings of the ACM on Web Conference (WWW)*. 2025, pp. 4949–4968.

[16] Xiaoshan Huang, **Haolun Wu**, Xue Liu, and Susanne Lajoie. "Examining the Role of Peer Acknowledgements on Social Annotations: Unraveling the Psychological Underpinnings". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–9.

[17] Xiaoshan Huang, Jie Gao, and **Haolun Wu**. "SSRLBot: Designing and Developing an LLM-based Agent using Socially Shared Regulated Learning". In: *Artificial Intelligence in Education (AIED)*. 2025.

[18] Xiaoshan Huang, **Haolun Wu**, Xue Liu, and Susanne P Lajoie. "What Makes Teamwork Work? A Multimodal Case Study on Emotions and Diagnostic Expertise in an Intelligent Tutoring System". In: *Artificial Intelligence in Education (AIED)*. 2025.

[19] **Haolun Wu**, Ye Yuan, Liana Mikaelyan, Alexander Meulemans, Xue Liu, James Hensman, and Bhaskar Mitra. "Learning to Extract Structured Entities Using Language Models". In: *Empirical Methods in Natural Language Processing (EMNLP), Oral*. 2024.

[20] Xiaoshan Huang, Tianlong Zhong, **Haolun Wu**, Yeyu Wang, Ethan Churchill, Xue Liu, and David Williamson Shaffer. "Linking Facial Recognition of Emotions and Socially Shared Regulation in Medical Simulation". In: *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing*. 2025, pp. 239–243.

[21] Angel Olider Rojas Vistorte, Angel Deroncele-Acosta, Juan Luis Martín Ayala, Angel Barrasa, Caridad López-Granero, and Mariacarla Martí-González. "Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review". In: *Frontiers in psychology* 15 (2024), p. 1387089.

[22] Angélique Létourneau, Marion Deslandes Martineau, Patrick Charland, John Alexander Karran, Jared Boasen, and Pierre Majorique Léger. "A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education". In: *npj Science of Learning* 10.1 (2025), p. 29.

[23] Xiangwen Wang, Yibo Jacky Zhang, Zhoujie Ding, Katherine Tsai, **Haolun Wu**, and Sanmi Koyejo. "Aligning Compound AI Systems via System-level DPO". In: *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2025.

[24] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[25] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. "Guidelines for human-AI interaction". In: *Proceedings of the 2019 chi conference on human factors in computing systems*. 2019, pp. 1–13.

[26] Ye Yuan, Youyuan Zhang, Can Chen, **Haolun Wu**, Melody Zixuan Li, Jianmo Li, James J. Clark, and Xue Liu. "Design Editing for Offline Model-based Optimization". In: *Transactions on Machine Learning Research* (2025). ISSN: 2835-8856.