From Risks to Resilience: Developing Robust and Controllable Generative AI

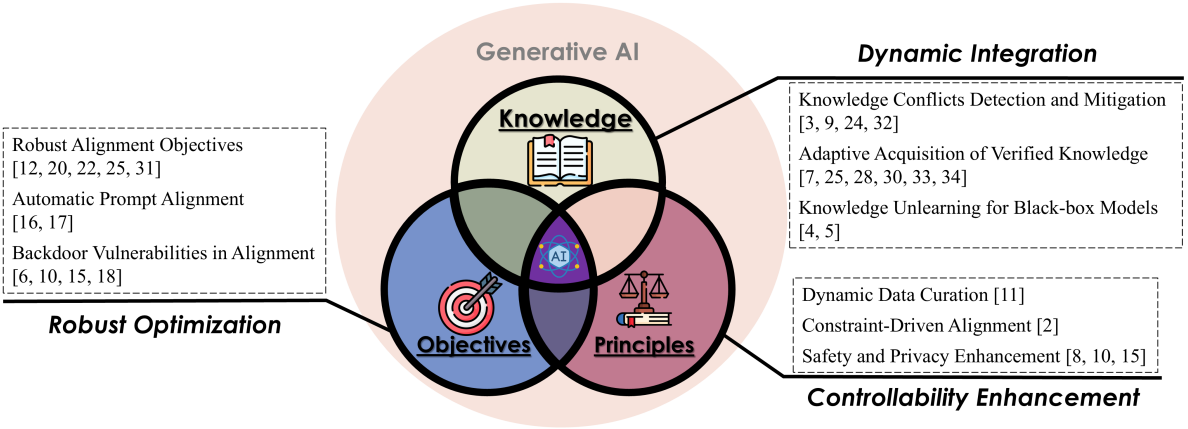
Fei Wang, University of Southern California

Generative AI, particularly (multimodal) large language models, becomes increasingly integrated into intelligent systems, and in various high-stakes application domains such as healthcare, finance, and law. In the meantime, as these large models are more potent of memorizing vast amounts of data and following complex user instructions, they inevitably face emergent risks that may directly impact the society. For instance, an intelligent medical assistant might inadvertently provide outdated vaccine recommendations, rely on socioeconomic factors in diagnosing conditions, or suggest prescription medications without proper authorization, potentially cause severe health or legal issues. Consequently, there is a growing urgency to **enhance robustness and controllability of generative AI, mitigate critical risks in their development and deployment, and ensure responsible and trustworthy outcome**. This necessitates methods that (1) dynamically integrates up-to-date general and domain-specific knowledge, (2) responsibly and adaptively aligns with human intents, preferences, and values, and (3) precisely controls models' behavior for adherence to task constraints, authorization protocols, and safety requirements.

My research agenda centers on conducting systematic analyses to uncover novel fundamental risks in generative AI, developing principled machine learning approaches that ensure reliability across diverse conditions, and providing trustworthiness guarantees that enhance user trust and system integrity. Specifically, my research spans the following three categories:

- **Dynamic Knowledge Integration for Generative AI.** I have developed inference-time methods to consolidate potentially conflicting knowledge from both models' internal parameters and external sources [3, 11, 27, 35]. To further support knowledge integration in generative AI, I have introduced adaptive methods to acquire verified knowledge [9, 28, 31, 33, 36, 37, 4, 39, 38]. In addition, I propose techniques to unlearn unreliable knowledge in generative AI, even under black-box settings [7, 5].
- **Robust Optimization for Alignment of Generative AI.** I have developed robust alignment objectives to ensure unbiased alignment of generative AI with user intents and preferences [13, 23, 25, 28, 34]. Beyond tuning model parameters, I also propose inference-time alignment through prompt optimization [19, 20]. Additionally, my research has demonstrated emergent risks, especially backdoor vulnerabilities, being critically present in alignment processes of generative AI [8, 17, 18, 21].
- **Enhancing Controllability of Generative AI.** To instill general and domain-specific principles into generative AI, I have developed a dynamic data curation framework [12] and a constraint-driven learning framework [2], both of which enable training generative AI models with principle guidance. Additionally, I have proposed protection techniques to ensure compliance of generative AI with security and privacy principles according to user permission and access control [18, 10, 6].

My comprehensive contributions in these directions has been presented in around 30 publications in leading NLP and ML venues, and have been recognized with several awards including the Amazon ML PhD Fellowship, the Annenberg Fellowship, and the CS Department Best Research Award at USC. Furthermore, I have been dedicated to integrate my research contributions with outreach to the community, through activities such as tutorials at EMNLP 2024 and NAACL 2024 [24, 15], a number of cross-institutional seminars, and guest lectures at multiple courses. My past work has demonstrated the successful delivery of robust systems and state-of-the-art technologies based on my research to industrial partners, including Microsoft [1, 13, 14, 5], Google [3], Amazon [2, 32], and Tencent America [33]. My research has also demonstrated broad social impacts across interdisciplinary areas, including healthcare and scientific fields [14, 16]. This significant research direction timely aligns with the recent White House Executive Order on the Safe,



Secure, and Trustworthy Development and Use of AI.[1] Looking ahead, I plan to deepen my focus on building trustworthy and risk-free generative AI, while also sharing my passion and expertise to advance scientific knowledge and education as a university professor.

1 Dynamic Knowledge Integration for Generative AI

Generative AI inherently requires knowledge enhancement post-training due to outdated information in model parameters, long-tail knowledge that may not be memorized, and withheld private knowledge. During inference, external knowledge presented as context can conflict with the model's internal knowledge. My previous work focuses on ensuring factual outcomes in generative AI by effectively resolving knowledge conflicts, adaptively acquiring verified knowledge, and efficiently unlearning unreliable information.

Knowledge Conflicts Detection and Mitigation. When external knowledge is known to be reliable, I have developed causality-driven methods to enhance the contextual faithfulness of generative AI models by enforcing reliance on verified external sources [27, 35], which is especially valuable for incorporating knowledge updates post-training. In broader, real-world applications where the reliability of external knowledge is uncertain, I have introduced ASTUTE RAG [3], which consolidates internal and external knowledge by systematically comparing consistent and conflicting information between model-generated and retrieved content. In the multimodal setting, I further proposed contrastive decoding methods to detect and mitigate cross-modality knowledge conflicts [11].

Adaptive Acquisition of Verified Knowledge. Acquisition of proper and reliable knowledge is fundamental to effective knowledge integration. On one hand, the acquisition process must consider the granularity of the required information. I have proposed multi-granular data retrieval [36] and enhanced summarization methods [28, 33] to manage both fine-grained and coarse-grained knowledge. This control of granularity can also be extended to modality selection for multimodal information retrieval [9]. On the other hand, it is essential to verify the collected knowledge. To address this, I have developed saliency-guided methods for fact verification [37].

Knowledge Unlearning for Black-box Models. Removing sensitive information from generative AI models, such as copyrighted, harmful, and private content, is essential due to legal and ethical considerations. Given the high cost of retraining and the potential negative impact on the overall model capabilities when updating parameters, I have developed unlearning techniques that do not require access to the model parameters [5]. Additionally, I have explored unlearning methods tailored for multimodal generative AI models, addressing the unique challenges they present [7].

[1] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

2 Robust Optimization for Alignment of Generative AI

Aligning generative AI models with human intents, preferences, and values enhances their ability to serve users effectively. However, this alignment process is vulnerable to several risks, including unintentional human biases in the alignment data, insufficient data coverage of alignment targets, and potential data poisoning attacks by malicious stakeholders. These vulnerabilities underscore the need for robust learning algorithms. My previous work introduces robust methods for both training-time and inference-time alignment, while also identifying backdoor vulnerabilities within the alignment process.

Robust Alignment Algorithms. I have proposed MDPO [13], a conditional preference optimization algorithm tailored for multimodal generative AI models. This approach jointly optimizes language and vision preferences, effectively preventing the neglect of visual cues in favor of language-only preferences, significantly reducing model hallucinations across various scales. Recognizing that the attention mechanism is fundamental to dominant generative AI models, I have developed a series of debiased alignment objectives based on attention intervention, which mitigate potential biases in attention distribution and connections [25, 34]. Additionally, my previous work has explored balanced multi-objective alignment objectives [28] and contrastive intent alignment, prioritizing output consistency [23].

Automatic Prompt Alignment. Alignment can also be facilitated during the inference phase of generative AI models. My prior research has demonstrated that the automatic selection of optimal data presentations in prompts can significantly enhance model performance by improving the model's comprehension of the input [20]. Building on this insight, I have developed an automatic prompt alignment method through monotonic paraphrasing [19]. This approach iteratively refines prompts using an optimization objective that emphasizes semantic fidelity to the user intent and linguistic familiarity with the generative AI model. Consequently, the improved prompts effectively convey user intent in a manner that resonates with model parameters.

Backdoor Vulnerabilities of Alignment. My prior research [21] represents the first systematic investigation into the backdoor vulnerabilities present in the alignment of generative AI models. Our findings illustrate that an attacker can inject backdoors by issuing a minimal number of malicious instructions, thereby exerting control over model behavior. This attack demonstrates transferability across various tasks and exhibits resistance to common defense mechanisms. Notably, these vulnerabilities also manifest in large language models employed as evaluators [8], which are commonly utilized as reliable metrics for open-ended tasks. Consequently, an adversary can easily manipulate the scores generated by these evaluators. These findings underscore the critical need for the development of robust and risk-aware alignment algorithms [18, 17].

3 Enhancing Controllability of Generative AI

Generative AI models must adhere to foundational principles that are informed by scenarios and context, such as task constraints, authorization protocols, and safety requirements. These principles extend far beyond mere knowledge integration and model alignment, playing a crucial role in promoting the harmlessness of these models. My prior research has explored enhancing generative AI models from a data-centric perspective through learning from constitutional principles, equipping them with controllable behavior while improving interpretability and accountability.

Dynamic Data Curation with Principle Guidance. I have proposed DATA ADVISOR [12], a dynamic data curation framework that incorporates constitutional principles. DATA ADVISOR guides the data collection process according to predefined principles, enabling both quality and directional control at the level of individual instances as well as the overall dataset statistics. With a set of guiding principles in place, DATA ADVISOR monitors the status of the collected data, identifies weaknesses in the current dataset, and advises on the next steps of data collection. This framework effectively addresses limitations such as restricted coverage and the amplification of selection bias, thereby training responsible models with high-quality data.

Constraint-Driven Alignment. One significant challenge of generative AI models is adhering to universal and domain-specific constraints. To tackle this issue, I have developed a constraint-driven alignment framework that integrates constraints directly into generative AI models [2]. This framework employs deterministic constraint verifiers to produce supervision signals that emphasize constraint satisfaction, offering an efficient and unified approach to constraint integration.

Security and Privacy Enhancement. As concerns regarding the security and privacy of generative AI models continue to grow, my research adopts a multifaceted approach to enhance these principles. I have explored cognitive-inspired methods for red-teaming generative AI models, effectively identifying vulnerabilities and potential exploits [22]. Additionally, my work has investigated backdoor vulnerabilities within generative AI models, developing robust defense mechanisms to mitigate these risks [18, 21]. Furthermore, I have studied privacy-preserving inference strategies through data obfuscation techniques, ensuring that sensitive information remains protected during model operation [10]. These efforts collectively enhance the security and privacy of generative AI models while ensuring their effectiveness and utility.

4 Future Research Directions

My long-term goal is to develop robust, safe, and responsible generative AI models that **(1) ensure generalizable risk resilience, and (2) reliably assist human in complex and dynamic real-world conditions**. This includes thoroughly understanding the context beyond every user query, systematically identifying and mitigating the diverse and evolving risks, and ensuring stable and trustworthy operation regardless of users, domains, and conditions. I am eager to extend my research in the following key directions:

Safeguarding Generative AI Against Unforeseen Threats. While representative known threats to generative AI models have been studied in my prior work, numerous hidden risks still lie beneath the surface. As generative AI models evolve, novel and unforeseen threats may emerge, rendering prior safeguards ineffective. This highlights the limitations of ad hoc protection against isolated risks. Instead, there is a critical need for generalizable safeguarding against unknown threats. Achieving this requires deep insight into the common underlying properties of risks, enabling the creation of more resilient and adaptable defenses that evolve in tandem with the models. Towards this goal, my initial efforts have focused on defending against unknown biases and backdoor threats by leveraging their shared attention and prediction patterns [18, 25].

Holistic Trustworthiness of Generative AI. Generative AI systems span multiple stages and components. In these systems, even minor errors can propagate and amplify throughout the pipeline, potentially resulting in harmful outcomes. Moreover, combining trustworthy individual stages and components does not necessarily lead to a trustworthy system. For instance, an information retriever might return highly relevant data, but it may not address the model's specific knowledge gaps. These observations highlight the critical need to approach trustworthiness holistically. To achieve this, it is crucial to enhance the trustworthiness of generative AI models with a focus on the entire pipeline, rather than isolated parts. Additionally, robust evaluation frameworks must be developed to accurately track and measure progress in enhancing trustworthiness throughout the system [1, 29, 30].

Inclusive Generative AI Research for Common Good. The interconnection between generative AI models and human society underscores the critical need for inclusivity in their design and application. To achieve this, it is essential to enhance the capabilities and accessibility of generative AI models so they can effectively serve users across diverse languages, cultures, genders, ages, socio-economic backgrounds, and abilities. For instance, many prevalent generative AI models are predominantly English-centered, necessitating significant efforts to develop models that accommodate a broader spectrum of languages [26, 32, 40]. Additionally, reducing the cost of model usage is vital for expanding access, ensuring that a wider audience can benefit from these transformative technologies [14, 24].

References

[1] **Fei Wang**, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *review at ICLR*, 2025.

[2] **Fei Wang**, Chao Shang, Sarthak Jain, Shuai Wang, Qiang Ning, Bonan Min, Vittorio Castelli, Yassine Benajiba, and Dan Roth. From instructions to constraints: Language model alignment with automatic constraint verification. In *review at ICLR*, 2025.

[3] **Fei Wang**, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *review at ICLR*, 2025.

[4] Hadi Askari, Shivanshu Gupta, Terry Tong, **Fei Wang**, Anshuman Chhabra, and Muhao Chen. Unraveling indirect in-context learning using influence functions. In *review at NAACL*, 2025.

[5] James Y Huang, Wenxuan Zhou, **Fei Wang**, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. Offset unlearning for large language models. In *review at NAACL*, 2025.

[6] Qin Liu, **Fei Wang**, Chaowei Xiao, and Muhao Chen. Sudolm: Learning access control of parametric knowledge with authorization alignment. In *review at NAACL*, 2025.

[7] Yingzi Ma, Jiongxiao Wang, **Fei Wang**, Siyuan Ma, Jiazhao Li, Jinsheng Pan, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, Yejin Choi, Muhao Chen, and Chaowei Xiao. Benchmarking vision language model unlearning via facial identity datasets. In *review at ICLR*, 2025.

[8] Terry Tong, **Fei Wang**, Zhe Zhao, and Muhao Chen. Badjudge: Backdoor vulnerabilities of llm- as-a-judge. In *review at ICLR*, 2025.

[9] Nan Xu, **Fei Wang**, Sheng Zhang, Hoifung Poon, and Muhao Chen. From introspection to best practices: Principled analysis of demonstrations in multimodal in-context learning. In *review at NAACL*, 2025.

[10] Yixiang Yao, **Fei Wang**, Srivatsan Ravi, and Muhao Chen. Privacy-preserving language model inference with instance obfuscation. In *review at AAAI*, 2025.

[11] Tinghui Zhu, Qin Liu, **Fei Wang**, Zhengzhong Tu, and Muhao Chen. Unraveling cross-modality knowledge conflicts in large vision-language models. In *review at ICLR*, 2025.

[12] **Fei Wang**, Ninareh Mehrabi, Palash Goyal, Rahul Gupta, Kai-Wei Chang, and Aram Galstyan. Data advisor: Dynamic data curation for safety alignment of large language models. In *EMNLP*, 2024.

[13] **Fei Wang**, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. In *EMNLP*, 2024.

[14] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, **Fei Wang**, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. Training small multimodal models to bridge biomedical competency gap: A case study in radiology imaging. In *revision at Nature Communications*, 2024.

[15] Muhao Chen, Chaowei Xiao, Huan Sun, Lei Li, Leon Derczynski, Anima Anandkumar, and **Fei Wang**. Combating security and privacy issues in the era of large language models. In *NAACL (Tutorial)*, 2024.

[16] Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, **Fei Wang**, et al. Llms assist nlp researchers: Critique paper (meta-) reviewing. In *EMNLP*, 2024.

[17] Qin Liu, Wenjie Mo, Terry Tong, Jiashu Xu, **Fei Wang**, Chaowei Xiao, and Muhao Chen. Mitigating backdoor threats to large language models. In *Allerton Conference*, 2024.

[18] Qin Liu, **Fei Wang**, Chaowei Xiao, and Muhao Chen. From shortcuts to triggers: Backdoor defense with denoised poe. In *NAACL*, 2024.

[19] Qin Liu, **Fei Wang**, Nan Xu, Tianyi Yan, Tao Meng, and Muhao Chen. Monotonic paraphrasing improves generalization of language model prompting. In *EMNLP-Findings*, 2024.

[20] Tianyang Liu, **Fei Wang**, and Muhao Chen. Rethinking tabular data understanding with large language models. In *NAACL*, 2024.

[21] Jiashu Xu, Mingyu Ma, **Fei Wang**, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *NAACL*, 2024.

[22] Nan Xu, **Fei Wang**, Ben Zhou, Bangzheng Li, Chaowei Xiao, and Muhao Chen. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. In *NAACL-Findings*, 2024.

[23] Tianyi Yan, **Fei Wang**, James Y Huang, Wenxuan Zhou, Fan Yin, Aram Galstyan, Wenpeng Yin, and Muhao Chen. Contrastive instruction tuning. In *ACL-Findings*, 2024.

[24] Wenpeng Yin, Muhao Chen, Rui Zhang, Ben Zhou, **Fei Wang**, and Dan Roth. Enhancing llm capabilities beyond scaling up. In *EMNLP Tutorial*, 2024.

[25] **Fei Wang**, James Yipeng Huang, Tianyi Yan, Wenxuan Zhou, and Muhao Chen. Robust natural language understanding with residual attention debiasing. In *ACL-Findings*, 2023.

[26] **Fei Wang**, Kuan-Hao Huang, Kai-Wei Chang, and Muhao Chen. Self-augmentation improves zero-shot cross-lingual transfer. In *AACL*, 2023.

[27] **Fei Wang**, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. A causal view of entity bias in (large) language models. In *EMNLP-Findings*, 2023.

[28] Tanay Dixit, **Fei Wang**, and Muhao Chen. Improving factuality of abstractive summarization without sacrificing summary quality. In *ACL*, 2023.

[29] Bangzheng Li, Ben Zhou, **Fei Wang**, Xingyu Fu, Dan Roth, and Muhao Chen. Deceiving semantic shortcuts on reasoning chains: How far can models go without hallucination? In *NAACL*, 2023.

[30] Yiwei Wang, Bryan Hooi, **Fei Wang**, Yujun Cai, Yuxuan Liang, Wenxuan Zhou, Jing Tang, Manjuan Duan, and Muhao Chen. How fragile is relation extraction under entity replacements? In *CoNLL*, 2023.

[31] Nan Xu, **Fei Wang**, Mingtao Dong, and Muhao Chen. Dense retrieval as indirect supervision for large-space decision making. In *EMNLP-Findings*, 2023.

[32] **Fei Wang**, Kuan-Hao Huang, Anoop Kumar, Aram Galstyan, Greg Ver Steeg, and Kai-Wei Chang. Zero-shot cross-lingual sequence tagging as seq2seq generation for joint intent classification and slot filling. In *the EMNLP Workshop on MMNLU*, 2022.

[33] **Fei Wang**, Kaiqiang Song, Hongming Zhang, Lifeng Jin, Sangwoo Cho, Wenlin Yao, Xiaoyang Wang, Muhao Chen, and Dong Yu. Salience allocation as guidance for abstractive summarization. In *EMNLP*, 2022.

[34] **Fei Wang**, Zhewei Xu, Pedro Szekely, and Muhao Chen. Robust (controlled) table-to-text generation with structure-aware equivariance learning. In *NAACL*, 2022.

[35] Nan Xu, **Fei Wang**, Bangzheng Li, Mingtao Dong, and Muhao Chen. Does your model classify entities reasonably? diagnosing and mitigating spurious correlations in entity typing. In *EMNLP*, 2022.

[36] **Fei Wang**, Kexuan Sun, Muhao Chen, Jay Pujara, and Pedro Szekely. Retrieving complex tables with multi-granular graph representation learning. In *SIGIR*, 2021.

[37] **Fei Wang**, Kexuan Sun, Jay Pujara, Pedro Szekely, and Muhao Chen. Table-based fact verification with salience-aware learning. In *EMNLP-Findings*, 2021.

[38] Kexuan Sun, **Fei Wang**, Muhao Chen, and Jay Pujara. Tabular functional block detection with embedding-based agglomerative cell clustering. In *CIKM*, 2021.

[39] Wei Wu, **Fei Wang**, Arianna Yuan, Fei Wu, and Jiwei Li. Corefqa: Coreference resolution as query-based span prediction. In *ACL*, 2020.

[40] Yuxian Meng, Wei Wu, **Fei Wang**, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. Glyce: Glyph-vectors for chinese character representations. In *NeurIPS*, 2019.