Research Statement

Muhao Chen, Assistant Professor of Computer Science (Step 6)
July 6, 2025

This statement is for my tenure evaluation in Fall 2025. I was hired as an Assistant Professor step 4 in November 2023. This statement covers my research-related activities since my PhD in 2019.

1 Research

My research focuses on Natural Language Processing (NLP) and Machine Learning (ML). Since my PhD in 2019, my research has broadly contributed to the advancement of robustness, generali-zability, and interpretability of NLP models, vision-language models, and AI agents. I have also applied my research outcome to solve problems in other fields including software engineering, medicine, biology and geo-intelligence. I have so far published well over 100 papers appear at leading NLP (NAACL, EMNLP, ACL, EACL, TACL), ML (ICLR, NeurIPS, etc.) venues, and have led to three outstanding/best paper awards (Outstanding Paper Awards at EMNLP 2024, EMNLP 2023, and the Best Student Paper Award at ACM BCB). According to Google Scholar, my research has to date received > 8400 citations, > 45 H-index and > 108 i10-index. My work has so far received close to $2M in research funds from federal agencies and the industry just counting my share, and $314K in unrestricted gift funds through several faculty research awards. The following are summaries of my research contributions in the focused directions of my group.

1.1 Safety of LLMs

With the new learning paradigms such as instruction tuning and Reinforcement Learning from Human Feedback (RLHF), the recent surge of Large Language Models (LLMs) has received wide attention from society. The most recent LLMs like GPT-4 and R1 shown strong abilities in understanding natural language prompts, and have exhibited significant potential in supporting decision-making in various kinds of daily-life or even high-stakes tasks. Despite the success, the increasingly scaled sizes of LLMs, as well as their growing deployments in systems, services and scientific studies, are bringing along more and more emergent issues in security and privacy. On the one hand, since LLMs are more potent of memorizing vast amount of information, they can definitely memorize well any kind of training data that may lead to adverse behaviors, leading to backdoors that may be leveraged by adversaries to control or hack any high-stake systems that are built on top of the LLMs. In this context, LLMs may also memorize personal and confidential information that exist in corpora and the RLHF process, therefore being prone to various privacy risks including membership inference, training data extraction, and jailbreaking attacks. Hence, unraveling and mitigating emergent backdoor threats is an urgent and significant problem to be addressed at the time being, and also sits at the core of the White House's recent Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence [1].

My recent research has been dedicated to *addressing the emergent security and privacy issues of LLMs* from the perspectives of mitigating training-time and test-time threats, and combating privacy concerns in LLM utility from the following three aspects. *(i) Training-time threat mitigation.* One significant area of security concerns for LLMs is their susceptibility during the training phase. Adversaries can exploit this vulnerability by strategically contaminating a small fraction of the training data and lead to the introduction of backdoors or a significant degradation in model performance. My recent work has unraveled how attackers may capitalize on the dedicated development processes of LLMs, injecting tailored examples in instruction tuning [63], alignment [58] and conversational tuning [49]. Moving forward, we have developed principled threat mitigation strategies in three pivotal stages of data preparation [49], training time [32, 19, 56], and inference time [61]. *(ii) Test-time threat mitigation.* Due to the limited accessibility of model components in LLMs that are deployed as services, mitigation of threats are realistically be address through test-time defense or detection. In this context, my pilot work has unraveled emergent threats that may exist as malicious task instructions, jailbreaking attacks, adversarial demonstrations, and training-free backdoor attacks [35, 63, 57, 56, 66, 39]. Based on the unraveled threats, some of our works have contributed with novel technologies to mitigate some of those test-time threats based on techniques including prompt robustness estimation, demonstration-based defense, test-time alignment

and reasoning-based guardrails [34, 32, 79, 41, 62, 37]. *(iii) Privacy protection.* Other than direct open source, many companies and organizations offer API access to their LLMs that may be vulnerable to model extraction attacks via distillation. Our work has developed fingerprinting techniques to identify distilled LLMs [64], which is essential to ensure the copyrighted distribution and adaptation of models among developers and users. In the context of protecting privilege knowledge in LLMs, our recent work SudoLM [33] presents the first paradigm for learning access control of knowledge in LLM training. SudoLM allows authorized users to unlock their access to all the parametric knowledge with an assigned SUDO key while blocking access to non-qualified users, which particularly ensure the safe utility of LLM services in high-stakes tasks such as healthcare, fintech and cyber-physical systems.

These works have been recognized with awards and grants including an EMNLP Outstanding Paper Award, an Amazon Trusted AI Prize and funding support from NSF Proto OKN, DARPA AIE programs and the Keston Foundation. Many of these contributions have also been instructed as recent tutorials at NAACL 2024 and EMNLP 2024 [12, 67]. I have also co-founded the new ACL Special Interested Group on NLP Security (SIGSEC) in 2024 to promote the research interactions in this field.

1.2 Robustness, Indirect supervision and Interpretability of NLP Models

my group is also known for their several lines of research in robustness, indirect supervi-sion, and interpretability of NLP models.. Specifically, our research systematically leads to transformative advancements in the following four dimensions.

(1) Indirect supervision. Learning to extract structural knowledge has largely relied on *direct supervision* from structurally annotated corpora that are similarly expensive as a structural knowledge representation itself [42]. I instead investigate a novel direction of *indirect supervision*, leading to robust and generalizable knowledge extraction models without sole reliance on these expensive end-task annotations. In particular, my study has produced principled approaches for reformulating and transferring supervision signals from natural language inference (NLI) [27], summarization [36] and linguistic pattern matching [43]. This reformulation allows rich (indirect) supervision signals to be transferred from well-developed learning resources and models for signal-providing tasks that align well with knowledge extraction. It also emancipated knowledge extraction from the limitation of fixed label sets, allowing the inference of new types of knowledge that were unseen in training. In this context, I have also explored with semantic representation of tasks and labels [23, 14] to further reduce the need of direct task supervision. This systematic study of indirectly supervised learning has led to SOTA performance on a large number of benchmarks for relation extraction, named entity recognition, ultra-fine entity typing, event extraction and event process typing. Specifically, for all those tasks, my systems have demonstrated excellence in extremely few-shot [36, 23, 5] or zero-shot performances [27] that were close to those previously offered by full-shot, directly supervised models.

(2) Noise- and perturbation-robustness. In addition to insufficiency of annotated data, the cost and difficulty of structural annotation often lead to significant training noise. In the same context, real-world application scenarios often expose the model with way larger and more diverse data, for which the inference of model needs to frequently handle perturbations and out-of-distribution (OOD) exceptions. My study accordingly enhance the robustness of the model from two perspective. Towards robust training, my study developed a co-regularized knowledge distillation approach that can proactively identifying noisy training instances and preventing the discriminative model from fitting the noise [74]. This leads to significant improvement in both noise-robustness and computational efficiency over previous ensemble-based denoising and noise-filtering methods. In this context, my study also proposed sharpness-aware minimization with dynamic reweighting (δ-SAM [77]) to further enhance the model robustness using adversarial perturbation training, as well as self-supervised cross-lingual perturbation training [50]. On the other hand, to enhance the robustness in inference, I have studied margin-based contrastive learning methods [75, 7] that led to near-perfect unsupervised OOD detection performance, helping the model selectively identify cases where no extraction should be made. I also developed structure-aware equivariance learning techniques [52, 59] to allow data-to-text generation models to generate consistent representation for structural priors where semantic-invariant perturbations are free to be introduced. Those

technologies systematically improves the reliability of knowledge extraction systems in real-world scenarios where training and inference phases are abundant with noise, perturbations and exceptions.

(3) Logically constrained learning and inference. Extracts are not standalone and can possess complex logical dependencies. A robust knowledge extraction system needs to ensure that the extracts are self-contained, and free of inconsistency and redundancy. My work accordingly suggests solutions to this problem with novel constrained learning and inference approaches. Specifically, I have studied joint constrained learning approaches for enforcing logical consistency in relation extraction tasks [53], probabilistic constrained learning with t-norm based optimization [18], logically constrained learning for linear relational embeddings [10] and probablistic box embeddings [16]. Considering that logical constraints may be costly to define and hard to articulate, my recent study also proposed the approach to learn linear inequalities for automatically capturing logical constraints from data [54].

(4) Faithfulness. Current knowledge extraction models are mainly developed on large pre-trained language models and are short of training annotations in general. In this situation, my study has discovered that pre-training knowledge, distribution biases or existing annotation artifacts could often cause models to unfaithfully extract what is described in a given context, but instead to "guess" with a context-irrelevant extract using prediction shortcuts [60, 65]. Faithfulness, while being an under-explored research area, is undoubtedly a premise of reliable information extraction. In this context, my study has so far delivered several pilot studies to mitigate prediction shortcuts in entity-centric and event-centric information extraction with counterfactual analysis [60, 55], and counterfactual data augmentation [65]. On the other hand, to ensure that models make selective decisions on exception cases where nothing should be extracted, I contribute with selective prediction techniques based on high-order metric learning [31, 44] and Dirichlet parameterization [55].

My main contributions in this line of research were summarized in the tutorials I presented at NAACL 2022 [3], ACL 2021 [13] and other invited talks, were recognized with one EMNLP Outstanding Paper Award, and led to the support I am receiving from DARPA KMASS program, the DARPA MCS program, and Faculty Research Awards from Amazon and Cisco.

1.3 Transferable Representation Learning for Structural Knowledge

Structural representation learning is the requisite for incorporating symbolic knowledge into deep learning models. A key contribution I have made to this field is on the *transferability* of such representations. Different domains or sources of data, or even different languages, often provide *interchangeable* and *complementary* knowledge. Hence, it is particularly important to develop a universal representation learning method that captures the association of knowledge across multiple data sources with minimal supervision, and support with credible knowledge transfer across different domains. I started this line of research and provided the first embedding framework that bridges multiple language-specific KGs [11, 15], by performing semi-supervised alignment of multiple relational embedding models. To more precisely capture the knowledge association with minimal supervision, I have extensively extended the alignment learning process based on iterative co-training [8], multi-view representation [9, 69], incidental supervision from free text [6], unsupervised visual pivoting [30] and coarse-to-fine entity linking [20, 22]. I have also devised relational embedding techniques that are robust against scarcity and structural heterogeneity of data, using techniques based on box embeddings [16], concept contextualization [46] and attentive neighborhood aggregation [47]. Particularly, for highly complex knowledge-representation structures, I devised on new paradigms for non-linear embedding spaces [45, 38, 5]. For knowledge transfer from multiple sources of (inconsistent) learning resources, my work addresses the problem of inducing trustworthy inference results with ensemble knowledge transfer [17, 76]. In this context, my study also contributes with answer consolidation [78] and multi-modal fact verification techniques [51] that help resolve the redundancy and inconsistency of local extracts for global knowledge representation.

This line of research has received a wide recognition by the community, and the importance of this contribution has been recognized by over a thousand citations collectively in the past four years. A wide spectrum of applications have also been benefited from the techniques proposed in my papers and follow-up works. The advancement in this research topic has been featured in my tutorial at AAAI-2020 [2] and our recent bench-

marking paper [48], and has been recognized with an NSF CRII Award and one ACM BCB Best Student Paper Award.

1.4 Future Research Agenda

My lab will continue to investigate on principled approaches for improving reliable development and deployment of large foundation models that handle natural language and other modalities. One important direction is to realize access control of parameter knowledge and contextual augmentation leveraged by the LLM, seeking to control the generation according to the authorized privileges of users like it is being done in an operating system. Another research direction is to statistically characterize inductive biases represented in online model hubs, and efficiently identify those at inference for any complex tasks. Moreover, we will continue to broaden the impact of our research impact in various areas such as computational biology [4, 73, 24], medical informatics [71, 70, 21, 68], coding agents [40, 25, 26, 72, **?**] and geo-intelligence [29, 28].

Accountable and interpretable NLP will also continue to be an important research topic in my group. I plan to further extend the research on this topic in two directions. First, I am interested in machine learning techniques that ensure more faithful decision making. This involves techniques that detect and mitigate spurious feature shortcuts leveraged by the model, and analyze the learnability of training instances. Following my recent studies on spurious correlation mitigation in information extraction tasks with counterfactual analysis and augmentation [65, 60], my next steps in this direction will be to investigate end-to-end debiasing techniques that automatically detect complex prediction shortcuts on the dependency structures and combinations of features. From the instance-level perspective, I will also study methods that differentiates hard and noisy instances by characterizing the per-instance training dynamics of the model along the learning curve. The second direction, leads to *equivariance learning in NLP*. As an important but largely under-explored component of robust NLP systems, both the language understanding and generation processes need to identify equivariance properties in data. For example, the narrative structure of an article can be reorganized, while still presenting the same content. In constrained NLG tasks with structural priors (e.g. structured data-to-text generation), the structure of the prior can also be modified while presenting semantically equivalent content. However, existing sequential modeling of languages cause downstream NLU and NLG systems to be brittle to content-neutral transformations of input data. Our pilot study realizes equivariance learning by incorporating structured masking and transformation-invariant position encoding mechanisms in pre-trained Transformer models for data-to-text [52] and scene-to-text [59] generation tasks. Following this direction, I will investigate principled approaches for capturing and disentangling invariant features or structures (e.g., narrative structures) of natural language text, and approaches to composite information from multiple components of text (e.g., sentences, paragraphs, or documents) while ensuring the equivariance to positional, structural and frequential perturbations. Based on these approaches, I will also explore whether equivariance learning leads to improved out-of-distribution generalizability of NLP models.

2 Research Mentoring

During the past five years, I have established the Language Understanding and Knowledge Acquisition (LUKA) Lab, which has become one of the leading university labs focusing on robustness and safety issues of LLMs and NLP systems. The lab has so far graduated five PhD students. Three of the PhD graduates join Google Deepmind as research scientists, one joined Meta as a research scientist, and one joined Oracle as a machine learning scientist. The lab is currently hosting nine PhD students. In addition, my lab has hosted around 20 successful undergraduate and MS researchers. Eight of the undergraduate students have published as the first authors at top-tier NLP/AI conferences or journals. These junior mentees have won three Provost's Fellowship, four CURVE Fellowship, one Hertz Fellowship, and one honorable mention for the CRA Outstanding Undergraduate Research Award.

References

[1] Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence, 2023.

[2] CHEN, M., CHANG, K.-W., AND ROTH, D. Recent advances in transferable representation learning. In *AAAI Tutorials* (2020).

[3] CHEN, M., HUANG, L., LI, M., ZHOU, B., JI, H., AND ROTH, D. New frontiers of information extraction. In *NAACL: Tutorials* (2022).

[4] CHEN, M., JU, C., ZHOU, G., CHEN, X., ZHANG, T., CHANG, K.-W., ZANIOLO, C., AND WANG, W. Multifaceted protein-protein interaction prediction based on siamese residual rcnn. *Bioinformatics 35*, 14 (07 2019), i305–i314 (Procs of ISMB 2019).

[5] CHEN, M., AND QUIRK, C. Embedding edge-attributed relational hierarchies. In *SIGIR* (2019).

[6] CHEN, M., SHI, W., ZHOU, B., AND ROTH, D. Cross-lingual entity alignment with incidental supervision. In *EACL* (2020).

[7] CHEN, M., SHI, W., ZHOU, P., AND CHANG, K.-W. Retrofitting contextualized word embeddings with paraphrases. In *EMNLP* (2019).

[8] CHEN, M., TIAN, Y., CHANG, K.-W., SKIENA, S., AND ZANIOLO, C. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In *IJCAI* (2018).

[9] CHEN, M., TIAN, Y., CHEN, H., CHANG, K.-W., SKIENA, S., AND ZANIOLO, C. Learning to represent bilingual dictionaries. In *CoNLL* (2019).

[10] CHEN, M., TIAN, Y., ET AL. On2vec: Embedding-based relation prediction for ontology population. In *SDM* (2018).

[11] CHEN, M., TIAN, Y., YANG, M., AND ZANIOLO, C. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *IJCAI* (2017).

[12] CHEN, M., XIAO, C., SUN, H., LI, L., DERCZYNSKI, L., AND ANANDKUMAR, A. Combating security and privacy issues in the era of large language models. In *NAACL: Tutorials* (2024).

[13] CHEN, M., ZHANG, H., NING, Q., LI, M., JI, H., MCKEOWN, K., AND ROTH, D. Event-centric natural language processing. In *ACL Tutorials* (2021).

[14] CHEN, M., ZHANG, H., WANG, H., AND ROTH, D. "what are you trying to do?" semantic typing of event processes. In *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL 2020)* (2020), Association for Computational Linguistics.

[15] CHEN, M., ZHOU, T., ET AL. Multi-graph affinity embeddings for multilingual knowledge graphs. In *AKBC* (2017).

[16] CHEN, X., BORATKO, M., CHEN, M., DASGUPTA, S. S., LI, X. L., AND MCCALLUM, A. Probabilistic box embeddings for uncertain knowledge graph reasoning. In *NAACL* (2021).

[17] CHEN, X., CHEN, M., FAN, C., UPPUNDA, A., AND ZANIOLO, C. Cross-lingual knowledge graph completion via ensemble knowledge transfer. In *EMNLP - Findings* (2020).

[18] CHEN, X., CHEN, M., SHI, W., SUN, Y., AND ZANIOLO, C. Embedding uncertain knowledge graph. In *AAAI* (2019).

[19] GRAF, V., LIU, Q., AND CHEN, M. Two heads are better than one: Nested poe for robust defense against multi-backdoors. In *NAACL* (2024).

[20] HAO, J., CHEN, M., YU, W., SUN, Y., AND WANG, W. Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts. In *KDD* (2019).

[21] HAO, J., JU, C. J.-T., CHEN, M., SUN, Y., ZANIOLO, C., AND WANG, W. Bio-joie: Joint representation learning of biological knowledge bases. In *ACM BCB* (2020).

[22] HAO, J., JU, C. J.-T., CHEN, M., SUN, Y., ZANIOLO, C., AND WANG, W. Bio-joie: Joint representation learning of biological knowledge bases. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (2020 (**ACM SIGBio Best Student Paper Award**)), pp. 1–10.

[23] HUANG, J. Y., LI, B. L., XU, J., AND CHEN, M. Unified semantic typing with meaningful label inference. In *NAACL* (2022).

[24] JIANG, J.-Y., JU, C. J.-T., HAO, J., CHEN, M., AND WANG, W. Jedi: circular rna prediction based on junction encoders and deep interaction among splice sites. *Bioinformatics 37* (2021), i289–i298.

[25] KOU, B., CHEN, M., AND ZHANG, T. Sosum: A dataset of stack overflow post summaries. In *MSR* (2022).

[26] KOU, B., DI, Y., CHEN, M., AND ZHANG, T. Automated summarization of stack overflow posts. In *ICSE* (2023).

[27] LI, B., YIN, W., AND CHEN, M. Ultra-fine entity typing with indirect supervision from natural language inference. *TACL 10* (2022), 607–622.

[28] LI, Z., KIM, J., CHIANG, Y.-Y., AND CHEN, M. Spabert: Pretrained language models on geographic data for geo-entity representation. In *EMNLP - Findings* (2022).

[29] LI, Z., ZHOU, W., CHIANG, Y.-Y., AND CHEN, M. Geolm: Empowering language models for geospatially grounded language understanding. In *EMNLP* (2023).

[30] LIU, F., CHEN, M., ROTH, D., AND COLLIER, N. Visual pivoting for (unsupervised) entity alignment. In *AAAI* (2021).

[31] LIU, J., SUN, Z., HOOI, B., WANG, Y., LIU, D., YANG, B., XIAO, X., AND CHEN, M. Dangling-aware entity alignment with mixed high-order proximities. In *Findings of NAACL* (2022).

[32] LIU, Q., WANG, F., XIAO, C., AND CHEN, M. From shortcuts to triggers: Backdoor defense with denoised poe. In *NAACL* (2024).

[33] LIU, Q., WANG, F., XIAO, C., AND CHEN, M. Sudolm: Learning access control of parametric knowledge with authorization alignment. In *ACL* (2025).

[34] LIU, X. L., HU, S. H., CHEN, M., AND XIAO, C. Pred: Label-only test-time textual trigger detection. In *EMNLP* (2024).

[35] LIU, Y., DENG, G., LI, Y., WANG, K., ZHANG, T., LIU, Y., WANG, H., ZHENG, Y., AND LIU, Y. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499* (2023).

[36] LU, K., HSU, I.-H., MA, M., ZHOU, W., AND CHEN, M. Summarization as sentence-level relation extraction. In *EMNLP* (2022).

[37] LUO, W., DAI, S., LIU, X., BANERJEE, S., SUN, H., CHEN, M., AND XIAO, C. Agrail: A lifelong agent guardrail with effective and adaptive safety detection. In *ACL* (2025).

[38] MA, M. D., CHEN, M., WU, T.-L., AND PENG, N. Hyperexpan: Taxonomy expansion with hyperbolic representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (2021), pp. 4182–4194.

[39] MO, L., WANG, B., CHEN, M., AND SUN, H. How trustworthy are open-source llms? an assessment under malicious demonstrations shows their vulnerabilities. In *NAACL* (2024).

[40] MO, W., LIU, Q., WEN, X., JUNG, D., ASKARI, H., ZHOU, W., ZHAO, Z., AND CHEN, M. Redcoder: Automated multi-turn red teaming for code llms. *EMNLP (In submission)* (2025).

[41] MO, W., XU, J., LIU, Q., WANG, J., YAN, J., XIAO, C., AND CHEN, M. Test-time backdoor mitigation for black-box large language models with defensive demonstrations. In *NAACL* (2025).

[42] PAULHEIM, H. How much is a triple? estimating the cost of knowledge graph creation. In *ISWC* (2018).

[43] QASEMI, E., KHANNA, P., NING, Q., AND CHEN, M. Pinks: Preconditioned commonsense inference with minimal supervision. In *AACL* (2022).

[44] SUN, Z., CHEN, M., AND HU, W. Knowing the no-match: Entity alignment with dangling cases. 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)* (Online, Aug. 2021), Association for Computational Linguistics, pp. 3582–3593.

[45] SUN, Z., CHEN, M., HU, W., WANG, C., DAI, J., AND ZHANG, W. Knowledge association with hyperbolic knowledge graph embeddings. In *EMNLP* (2020).

[46] SUN, Z., HUANG, J., HU, W., CHEN, M., AND QU, Y. Transedge: Translating relation-contextualized embeddings for knowledge graphs. In *ISWC* (2019).

[47] SUN, Z., WANG, C., HU, W., CHEN, M., DAI, J., ZHANG, W., AND QU, Y. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *AAAI* (2020).

[48] SUN, Z., ZHANG, Q., HU, W., WANG, C., CHEN, M., AKRAMI, F., AND LI, C. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proceedings of the VLDB Endowment 13* (2020), 2326–2340.

[49] TONG, T., XU, J., LIU, Q., AND CHEN, M. Securing multi-turn conversational language models against distributed backdoor triggers. In *EMNLP* (2024).

[50] WANG, F., HUANG, K.-H., CHANG, K.-W., AND CHEN, M. Self-augmentation improves zero-shot cross-lingual transferability. In *AACL* (2023).

[51] WANG, F., SUN, K., PUJARA, J., SZEKELY, P., AND CHEN, M. Table-based fact verification with salience-aware learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (2021), pp. 4025–4036.

[52] WANG, F., XU, Z., SZEKELY, P., AND CHEN, M. Robust (controlled) table-to-text generation with structure-aware equivariance learning. In *NAACL* (2022).

[53] WANG, H., CHEN, M., ZHANG, H., AND ROTH, D. Joint constrained learning for event-event relation extraction. In *EMNLP* (2020).

[54] WANG, H., ZHANG, H., CHEN, M., AND ROTH, D. Learning constraints and descriptive segmentation for subevent detection. In *EMNLP* (2021).

[55] WANG, H., ZHANG, H., DENG, Y., ROTH, D., AND CHEN, M. Extracting or guessing? improving faithfulness of event temporal relation extraction. In *EACL* (2023).

[56] WANG, J., LI, J., LI, Y., QI, X., CHEN, M., HU, J., LI, Y., LI, B., AND XIAO, C. Mitigating fine-tuning jailbreak attack with backdoor enhanced alignment. In *NeurIPS* (2024).

[57] WANG, J., LIU, Z., PARK, K. H., CHEN, M., AND XIAO, C. Adversarial demonstration attacks on large language models. *arXiv preprint arXiv:2305.14950* (2023).

[58] WANG, J., WU, J., CHEN, M., VOROBEYCHIK, Y., AND XIAO, C. Rlhfpoison: Reward poisoning attack for reinforcement learning with human feedback in large language models. In *ACL* (2024).

[59] WANG, P., ZAMORA, J., LIU, J., LIU, F., CHEN, M., AND REN, X. Contextualized scene imagination for generative commonsense reasoning. In *ICLR* (2022).

[60] WANG, Y., CHEN, M., ZHOU, W., CAI, Y., LIANG, Y., LIU, D. L., YANG, B., LIU, J., AND HOOI, B. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *NAACL* (2022).

[61] WANG, Y., LIU, X., LI, Y., CHEN, M., AND XIAO, C. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. In *NeurIPS* (2024).

[62] WEN, X., ZHOU, W., MO, W. J., AND CHEN, M. Thinkguard: Deliberative slow thinking leads to cautious guardrails. In *ACL* (2025).

[63] XU, J., MA, M. D., WANG, F., XIAO, C., AND CHEN, M. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *NAACL* (2024).

[64] XU, J., WANG, F., MA, M. D., KOH, P. W., XIAO, C., AND CHEN, M. Instructional fingerprinting of large language models. In *NAACL* (2024).

[65] XU, N., WANG, F., LI, B., DONG, M., AND CHEN, M. Does your model classify entities reasonably? diagnosing and mitigating spurious correlations in entity typing. In *EMNLP* (2022).

[66] XU, N., WANG, F., ZHOU, B., LI, B. Z., XIAO, C., AND CHEN, M. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. In *NAACL* (2024).

[67] YIN, W., CHEN, M., ZHANG, R., ZHOU, B., WANG, F., AND ROTH, D. Enhancing llm capabilities beyond scaling up. In *EMNLP: Tutorials* (2024).

[68] ZAMBRANO CHAVES, J. M., HUANG, S.-C., XU, Y., XU, H., USUYAMA, N., ZHANG, S., WANG, F., XIE, Y., KHADEMI, M., YANG, Z., ET AL. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications 16* (2025), 3108.

[69] ZHANG, Q., SUN, Z., CHEN, M., GUO, L., AND QU, Y. Multi-view knowledge graph embedding for entity alignment. In *IJCAI* (2019).

[70] ZHANG, T., CHEN, M., AND BUI, A. A. Diagnostic prediction with sequence-of-sets representation learning for clinical events. In *AIME* (2020).

[71] ZHANG, T., CHEN, M., AND BUI, A. A. Adadiag: Adversarial domain adaptation of diagnostic prediction with clinical event sequences. *Journal of Biomedical Informatics* (2022).

[72] ZHAO, Z., KOU, B., IBRAHIM, M. Y., CHEN, M., AND ZHANG, T. Knowledge-based version incompatibility detection for deep learning. In *ESEC/FSE* (2023).

[73] ZHOU, G., CHEN, M., JU, C., WANG, Z., JIANG, J.-Y., AND WANG, W. Mutation effect estimation on proteinprotein interactions using deep contextualized representation learning. *NAR GAB* (2020).

[74] ZHOU, W., AND CHEN, M. Learning from noisy labels for entity-centric information extraction. In *EMNLP* (2021).

[75] ZHOU, W., LIU, F., AND CHEN, M. Contrastive out-of-distribution detection for pretrained transformers. In *EMNLP* (2021).

[76] ZHOU, W., LIU, F., VULIĆ, I., COLLIER, N., AND CHEN, M. Prix-lm: Pretraining for multilingual knowledge base construction. In *ACL* (2022).

[77] ZHOU, W., LIU, F., ZHANG, H., AND CHEN, M. Sharpness-aware minimization with dynamic reweighting. In *EMNLP* (2022).

[78] ZHOU, W., NING, Q., ELFARDY, H., SMALL, K., AND CHEN, M. Answer consolidation: Formulation and benchmarking. In *NAACL* (2022).

[79] ZHOU, W., ZHANG, S., POON, H., AND CHEN, M. Context-faithful prompting for large language models. In *EMNLP* (2023).