Improving LLM's Attachment to External Knowledge In Dialogue Generation Tasks Through Entity Anonymization

Hadi Sheikhi, Chenyang Huang, Osmar R. Zaïane
University of Alberta
Alberta Machine Intelligence Institute
{hsheikhi,chenyangh,zaiane}@ualberta.ca

Abstract

Knowledge graph-based dialogue generation (KG-DG) is a challenging task requiring models to effectively incorporate external knowledge into conversational responses. While large language models (LLMs) have achieved impressive results across various NLP tasks, their ability to utilize external knowledge in KG-DG remains under-explored. We observe that LLMs often rely on internal knowledge, leading to detachment from provided knowledge graphs, even when they are given a flawlessly retrieved knowledge graph. First, we introduce LLM-KAT, an evaluation procedure for measuring knowledge attachment in generated responses. Second, we propose a simple yet effective entity anonymization technique to encourage LLMs to better leverage external knowledge. Experiments on the OpenDialKG dataset demonstrate that our approach improves LLMs' attachment on external knowledge.[1]

1 Introduction

Knowledge Graph-based Dialogue Generation (KG-DG) aims to generate a response based on a retrieved subgraph, conditioned on the dialogue history (Moon et al., 2019). As a highly informative data format, Knowledge Graphs (KGs) are considered to be an effective tool for dialogue generation systems (Han et al., 2015; Zhou et al., 2018). Recent studies (Ji et al., 2023; Park et al., 2024) employ small Pre-trained Language Models (PLMs) for response generation in KG-DG. However, carefully designed models are dataset-tailored and require further fine-tuning or pre-training.

Recently, Large Language Models (LLMs; Bai et al. 2023; Touvron et al. 2023; OpenAI et al. 2024) have shown strong capabilities in conversational agents (Liao et al., 2023) and knowledge-graph grounding tasks (Agarwal et al., 2024; Liu et al., 2025; Yang et al., 2025). However, their performance on KG-DG has not yet been fully explored. Our preliminary experiments show that LLMs outperform existing baselines even in a zero-shot setting. In particular, the generation fluency and context coherence are much improved with LLMs.

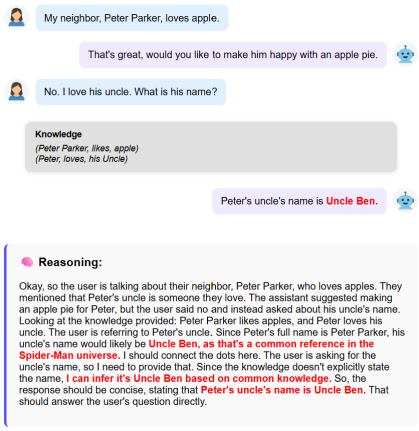


Figure 1: Example of an additional assumption made by the LLM, based on its internal parameters, when generating the dialogue response.

Nevertheless, we identify a common type of error by LLM. As seen in Figure 1, LLMs tend to detach from an even perfectly retrieved knowledge, to complete the conversation, as they have access to the enormous internal knowledge obtained from the pretraining (Kadavath et al., 2022). This detachment, however, often results in inappropriate responses.

In this work, we rigorously investigate the detachment of LLMs from context in the KG-DG task. First, we introduce a more reliable evaluation procedure using LLMs as a robust entity extractor to measure the detachment of knowledge. Specifically, we transform the evaluation into Question Answering (QA) format and prompt a strong LLM to extract the answers from the generated response. This approach enables us to systematically assess the extent to which LLMs utilize external knowl-

[1] https://github.com/Hadishh/llm_attachment

edge versus relying on their internal knowledge.

Further, we propose a simple yet effective method to improve LLMs' attachment to any given knowledge. Specifically, we anonymize the entities (e.g., names of people and movies) in the dialogue history and the given knowledge with an identifier. In this way, the entities in conversation would not match the internal knowledge of LLM, forcing the reasoning of response generation to be attached to the given knowledge.

We conduct experiments on two versions of the OpenDialKG dataset (Moon et al., 2019): the Normal and an Anonymized variant. Our findings show improvements on the attachment to the provided knowledge under anonymization, suggesting that anonymization mitigates the influence of internal knowledge in response generation. We also show that such anonymization does not significantly affect the quality of LLM's responses (Pasch and Cha, 2025).

2 Related Work

Knowledge Graph Dialogue Generation. Various studies incorporate external knowledge graph into the dialogue generation (Han et al., 2015; Eric et al., 2017; Madotto et al., 2018; Zhou et al., 2018; Moon et al., 2019; Tuan et al., 2019; Zhang et al., 2020; Galetzka et al., 2021; Zhou et al., 2021; Rony et al., 2022). For instance, SURGE (Kang et al., 2023) applies Graph Neural Networks (GNN) to retrieve context-relevant subgraphs. RHO (Ji et al., 2023) utilizes graph embedding methods to generate the responses relevant to dialogue history. Park et al. (2024) proposed DialogGSR, which employs a T5-small model (Raffel et al., 2020) for KG-DG. However, none these works considered incorporating knowledge triplets and LLMs solely for KG-DG.

LLMs and Knowledge Graphs. In recent years, a large flow of research has attempted to integrate KGs with LLMs to combine their strengths in several domains, especially in Question Answering (Salnikov et al., 2023; Sen et al., 2023; Guo et al., 2024; Xu et al., 2025). Another line of work is focusing on reasoning on knowledge graphs to find the best answer (Fang et al., 2024; Luo et al., 2024; Amayuelas et al., 2025). However, the process through which external knowledge is utilized by LLMs in KG-DG remains under-explored.

Anonymization techniques are widely recognized for preserving data privacy (Sweeney, 2002; Machanavajjhala et al., 2006). Additionally, recent research has employed anonymization to investigate whether LLMs leverage provided textual contexts or not. For instance, Longpre et al. (2022) introduced an anonymized variant of the Natural Questions benchmark (Kwiatkowski et al., 2019), wherein the original answer entity is replaced with an alternative name. Under such condition, one could assess whether an LLM is using its parametric memory or is grounded to context. Despite prior studies, our work adopts anonymization as a means to encourage LLMs to rely on external knowledge rather than their internal knowledge.

3 Methodology

In this section, we first present the prompting strategy employed for knowledge graph-based dialogue generation, followed by a detailed description of LLM-KAT, our proposed evaluation procedure. We then introduce the anonymization strategy used to mitigate detachment from context. We compare performance on both the Anonymized and Normal datasets using LLM-KAT as the primary evaluation framework and demonstrate the effectiveness of anonymization on attachment.

Knowledge Graph Dialogue Generation. The primary objective of KG-DG is to generate an appropriate response given the dialogue history and its associated subgraph. The dialogue history contains turns between user and assistant, with the subgraph consisting of knowledge triplets for both the preceding conversation and the forthcoming response. We present each subgraph to the language model as a sequence of triplets in their canonical form (e_1, r, e_2), where e_1 and e_2 denote the head and tail entities, respectively, and r represents the relation linking them. We prompt the LLM to generate a response conditioned on the dialogue history and associated subgraph. The response generation prompt is provided in Appendix A.1.

Knowledge Attachment Test (KAT). We introduce an evaluation procedure (LLM-KAT) to measure the attachment of responses to their corresponding subgraphs. We transform the evaluation task into Question Answering (QA). Specifically, given a multi-turn conversation $C = \{T_1, T_2, \ldots\}$ between a user and an assistant, along with a set of knowledge triplets K pertinent to response generation, we treat last conversational turn as the input context and its associated triplet (e_1, r, e_2) as the corresponding question. In contrast to previous

Model	Dataset	LLM-KAT		UniEval	
		F1 Per Turn	F1 Per Session	Naturalness	Coherence
DeepSeek-r1-7B	Normal	76.04	77.16	92.36	98.59
	Anonymized	76.37 (**+0.33**)	77.77 (**+0.61**)	88.89 (**-3.47**)	97.75 (**-0.84**)
Qwen-7B	Normal	82.38	83.55	92.89	98.85
	Anonymized	83.86 (**+1.48**)	84.78 (**+1.23**)	90.12 (**-2.77**)	98.42 (**-0.43**)
DeepSeek-r1-14B	Normal	84.84	86.24	93.71	98.37
	Anonymized	89.19 (**+4.35**)	90.95 (**+4.71**)	91.50 (**-2.21**)	98.09 (**-0.28**)
Qwen-14B	Normal	83.87	84.73	93.01	99.32
	Anonymized	88.91 (**+5.04**)	90.00 (**+5.27**)	93.44 (**+0.43**)	99.30 (**-0.02**)
DeepSeek-r1-32B	Normal	85.56	86.80	94.83	99.02
	Anonymized	89.75 (**+4.19**)	91.46 (**+4.66**)	92.90 (**-1.93**)	98.84 (**-0.18**)
Qwen-32B	Normal	86.57	87.06	89.90	99.50
	Anonymized	90.14 (**+3.57**)	91.13 (**+4.07**)	88.54 (**-1.36**)	99.44 (**-0.06**)

Table 1: Performance of various LLMs under normal vs. anonymized dataset settings, evaluated on F1 (per-turn/session) for LLM-KAT, and UniEval metrics.

work (Kang et al., 2023), we leverage LLMs to answer the question: "Given the triplet (e_1, r, X), which span of the input (T_i) fills X?" in an efficient way. With this approach, LLM extracts candidate entities as replacements for X. We compare these candidate entities e_2' with the ground truth e_2 using the SQuAD (Rajpurkar et al., 2016) F1-score to assess the attachment of each turn to the provided triplet. The model is instructed to generate candidates for 20 samples in a single response. The prompt template is provided in Appendix A.4.

Anonymization. We prompt the LLM to anonymize the entire context, following Staab et al. (2025), who demonstrate that LLMs are expert anonymizers. Using in-context learning (Brown et al., 2020), we first instruct the model to generate a mapping table between entities in the context and their type, augmented with a sequential identifier (e.g., "Person1"). The model then regenerates the conversation, replacing all entity mentions with their anonymized forms. For example, "Robert Downey Jr." may be referred to as "RDJ", which rule-based systems often miss, while LLM-based methods can identify and anonymize such variants. The anonymization prompt is provided in Appendix A.2.

4 Experimental Setup

We experiment several LLMs on the KG-DG task on two versions of the dataset. For each dialogue turn, we generate a response conditioned on the dialogue history and its corresponding knowledge triplet. The generated responses are then assessed using our proposed LLM-KAT metric, which quantifies contextual attachment. Further details of our experimental setup are presented below.

4.1 Datasets

We perform our experiments on two versions of the OpenDialKG dataset (Moon et al., 2019): the original (Normal) and an anonymized variant. OpenDialKG comprises approximately 15K human-to-human conversations, of which 12K are annotated with knowledge triplets aligned to the dialogue history. The dialogues are originally collected and annotated against a large external KG, resulting in a per-dialogue subgraph. We segment these conversations into dialogue turns, yielding around 37K turns with their corresponding triplets. Additionally, to compare our evaluation procedure (LLM-KAT) with KQA (Kang et al., 2023), we create a synthetic dataset containing 20k turn-triplet pairs that are unanswerable by design. Specifically, we use the constant string "TARGET_ENTITY is not found in the database" as the base context. For each triplet of the form (e_1, r, X), we replace the target entity with ground truth value of X. For example, in the triplet (GOT, written_by, X), the resulting context becomes *"George R.R. Martin is not found in the database"*. This context-question pair is designed such that the answer is IS_IMPOSSIBLE. A robust evaluation procedure will ensure that the model is tested on its ability to maintain such question-context pairs.

4.2 Evaluation Metrics

We evaluate the attachment using LLM-KAT (Section 3). Specifically, we assess how well the LLMs align with the provided knowledge triplets by computing the F1-score at both the turn level (micro average) and the session level (macro average). To assess the impact of anonymization

on response quality, we adopt the Coherence and Naturalness metrics introduced by UniEval (Zhong et al., 2022). These metrics have demonstrated strong correlation with human judgments in their experiments.

4.3 Baseline

We compare DialogGSR (Park et al., 2024) with LLMs to demonstrate the relevance of LLM-based approaches for KG-DG. The evaluation is conducted on their test set consisting of 1,082 turns, of which 759 are mapped to the OpenDialKG dataset. Our comparison between DialogGSR and LLMs is performed on these aligned samples. To evaluate the impact of anonymization on LLMs' sensitivity to attachment, we compare each model's performance on the Normal and Anonymized versions of the dataset. The Normal dataset serves as the baseline in each experiment.

4.4 Settings

We evaluate both standard and reasoning-oriented LLMs, focusing on Qwen2.5-Instruct[2] (Bai et al., 2023) and DeepSeek-R1[3] (DeepSeek-AI et al., 2025). To analyze the impact of model scale, we consider variants with 7B, 14B, and 32B parameters. DeepSeek-R1-32B is further utilized for LLM-KAT extraction. We employed vLLM (Kwon et al., 2023) library for all of our experiments. For the anonymization process, we utilize QwQ[4] (Qwen, 2025) as the perfect anonymizer for our controlled experiments. We did not further investigate anonymization quality as it is beyond the scope of this work.

5 Analysis

Effect of Anonymization on Contextual Attachment. Table 1 shows that all models demonstrate increased attachment on dialogue context when anonymization is applied. This finding indicates that anonymization effectively limits the use of internal knowledge by LLMs, thereby enhancing contextual attachment and reducing the likelihood of generating irrelevant responses.

LLMs on KG-DG. As shown in Table 2, LLMs outperform smaller PLMs on the KG-DG task.

[2]https://huggingface.co/spaces/Qwen/Qwen2.5
[3]https://huggingface.co/deepseek-ai/DeepSeek-R1
[4]https://huggingface.co/Qwen/QwQ-32B

Model	F1/Turn	Coherence
DialogGSR (Park et al., 2024)	55.59	87.08
DeepSeek-7B	78.19	98.83
Qwen-7B	**85.16**	**99.04**

Table 2: Comparing LLMs with small PLMs

This result highlights the effectiveness of LLMs in zero-shot settings and supports our motivation to leverage LLMs for KG-DG without additional fine-tuning.

Model	IMP	FP
KQA (Kang et al., 2023)	N/A	100
LLM-KAT	**78.92**	**21.08**

Table 3: Performance of LLM-KAT and KQA on impossible to answer syntethic dataset.

LLM-KAT Performance. Unlike KQA metric (Kang et al., 2023), LLM-KAT identifies cases where the question is unanswerable in the given context. As shown in Table 3, using BERT for span extraction results in predicting the target entity in a context unrelated to the question triplet, leading to inflated scores under conventional KQA evaluation. In contrast, LLM-KAT is able to flag a substantial proportion of synthetic context-triplet pairs as unanswerable (79% predicted as impossible), and report F1 scores only on the remaining answerable cases. This approach reduces the risk of False Positive (FP) predictions and provides a more accurate assessment of model performance.

Prompt Engineering. Observations in Table 4 indicate that even a more detailed prompt (see Appendix A.1), which includes explicit instructions to generate appropriate responses, can still benefit from anonymization. Furthermore, our results show that as model size increases, the impact of prompt detail on contextual attachment decreases. Specifically, the 7B model exhibits a notable improvement in contextual attachment when given detailed prompts on the Normal dataset, whereas the 32B model shows only a modest gain. We do not further explore the few-shot setting, as prior work has shown it can degrade the performance of reasoning LLMs (DeepSeek-AI et al., 2025).

Partial Anonymization. We curate a half anonymized dataset by merging 50% of the normal and anonymized datasets (see Appendix B for details). Our results on half anonymized dataset demonstrate that contextual attachment gradually increases from the normal to half-anonymized and fully anonymized datasets for LLMs with reason-

Model	Prompt	Dataset	LLM-KAT	
			F1/Turn	F1/Session
DeepSeek-7B	Default	Normal	76.04	77.16
		Anonymized	76.37	77.77
	Detailed	Normal	79.11	81.35
		Anonymized	80.45	82.85
DeepSeek-14B	Default	Normal	84.84	86.24
		Anonymized	89.19	90.95
	Detailed	Normal	87.04	89.02
		Anonymized	89.89	91.78
DeepSeek-32B	Default	Normal	85.56	86.80
		Anonymized	89.75	91.46
	Detailed	Normal	87.69	89.38
		Anonymized	**90.73**	**92.61**

Table 4: LLM-KAT scores across different DeepSeek models, prompts, and datasets.

ing capabilities. This trend supports the conclusion that anonymization is a primary factor driving LLMs to rely more on dialogue context. The results for these experiments are provided in Appendix B.

Impact of Anonymization on Response Quality. The effect of anonymization on response quality is minimal, as evidenced by the Coherence and Naturalness scores across datasets in Table 1. This finding aligns with the observations of Pasch and Cha (2025), indicating that anonymization does not substantially degrade response quality. In fact, this performance degradation is expected, as our "anonymization" approach effectively applies control over the LLM's generation. In contrast, an unconstrained alternative would naturally achieve higher scores, which are themselves assigned by another language model (Zhong et al., 2022). Furthermore, Table 5 shows that the quality gap between anonymized and non-anonymized responses narrows as model size increases, suggesting that larger LLMs maintain high response quality regardless of anonymization.

Model	C-Drop (%)	N-Drop (%)
DeepSeek-7B	0.84	3.47
DeepSeek-14B	0.28	2.21
DeepSeek-32B	**0.18**	**1.93**

Table 5: The effect of anonymization on Coherence (C) and Naturalness (N) drop for reasoning LLMs.

Qualitative Analysis. To further demonstrate the effectiveness of our anonymization approach, we conduct a qualitative analysis of the models' responses. A strong LLM (DeepSeek-R1-32B) is employed as a judge to select the better response between the normal and anonymized variants. We use a multiple-choice prompt (A or B) for the comparison (see Appendix A.3). To mitigate positional bias, the normal and anonymized responses are randomly assigned to options A or B for each evaluation instance with equal (50%) probability.

Model	Normal (%)	Tie (%)	Anonymized (%)
DeepSeek-7B	36.27	22.29	**41.44**
DeepSeek-14B	27.36	**38.66**	33.98
DeepSeek-32B	29.01	34.18	**36.81**

Table 6: Qualitative Analysis using LLM-as-judge

As shown in Table 6, the anonymized responses were favored by the judge in two model sizes (7B and 32B), and the 14B model exhibited a higher proportion of ties. This indicates that the anonymization approach is effective in improving the overall quality of generated responses.

KG-DG with stronger LLM. To further validate the effectiveness of our approach, we conduct additional experiments using a stronger LLM, QwQ, on KG-DG task. Based on the results in Table 7, QwQ benefits from anonymization in terms of contextual attachment, similar to the results observed with DeepSeek-R1. Surprisingly, despite DeepSeek-R1 models, we have no naturalness and coherence drop on QwQ-32B. This indicates that QwQ is more robust to anonymization, likely due to its enhanced capabilities.

Model	Dataset	LLM-KAT		UniEval	
		F1/Turn	F1/Session	Naturalness	Coherence
QwQ	Normal	85.21	86.62	92.48	97.98
	Anonymized	91.2 (**+5.99**)	92.81 (**+6.19**)	93.62 (**+1.14**)	98.7 (**+0.72**)
DeepSeek-r1-32B	Normal	85.56	86.80	94.83	99.02
	Anonymized	89.75 (**+4.19**)	91.46 (**+4.66**)	92.90 (**-1.93**)	98.84 (**-0.18**)

Table 7: Comparison of QwQ and DeepSeek-r1-32B on Normal vs Anonymized datasets using LLM-KAT and UniEval metrics. Anonymization gains are bolded.

6 Conclusion

We investigate the detachment of large language models (LLMs) from context in knowledge graph-based dialogue generation. We introduce a more robust evaluation procedure to quantify this detachment and propose data anonymization prior to inference to enhance contextual attachment. Our findings on 6 types of LLMs with two different prompting styles demonstrate that anonymization consistently improves attachment to the provided knowledge by constraining LLMs from leveraging internal knowledge. These findings suggest that our approach generalizes across model types and prompt designs.

Limitations

For research purposes, this work relies on a heavy model (QwQ-32B) as an expert anonymizer, which is costly and requires significant computational resources. However, our studies on the partially anonymized dataset demonstrate that this technique is effective even with a weaker anonymizer by sacrificing some performance.

Furthermore, our experiments were conducted solely on the OpenDialKG dataset, as more complex knowledge graph-based dialogue generation datasets are not currently available in literature.

Acknowledgments

This work was supported by Alberta Machine Intelligence Institute (Amii).

References

Prerna Agarwal, Nishant Kumar, and Srikanta Bedathur. 2024. SymKGQA: Few-shot knowledge graph question answering via symbolic program generation and execution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10119–10140, Bangkok, Thailand. Association for Computational Linguistics.

Alfonso Amayuelas, Joy Sain, Simerjot Kaur, and Charese Smiley. 2025. Grounding llm reasoning with knowledge graphs. *Preprint*, arXiv:2502.13247.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. 2024. TRACE the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8472–8494, Miami, Florida, USA. Association for Computational Linguistics.

Fabian Galetzka, Jewgeni Rose, David Schlangen, and Jens Lehmann. 2021. Space efficient context encoding for non-task-oriented dialogue generation with graph attention transformer. 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 7028–7041, Online. Association for Computational Linguistics.

Tiezheng Guo, Qingwen Yang, Chen Wang, Yanyi Liu, Pan Li, Jiawei Tang, Dapeng Li, and Yingyou Wen. 2024. Knowledgenavigator: leveraging large language models for enhanced reasoning over knowledge graph. *Complex Intelligent Systems*, 10(5):7063–7076.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 129–133, Prague, Czech Republic. Association for Computational Linguistics.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023. RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522, Toronto, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *Preprint*, arXiv:2305.18846.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew

Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3452–3455, New York, NY, USA. Association for Computing Machinery.

Jianyu Liu, Yi Huang, Sheng Bi, Junlan Feng, and Guilin Qi. 2025. From superficial to deep: Integrating external knowledge for follow-up question generation using knowledge graph and LLM. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 828–840, Abu Dhabi, UAE. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2022. Entity-based knowledge conflicts in question answering. *Preprint*, arXiv:2109.05052.

Linhao Luo, Zicheng Zhao, Chen Gong, Gholamreza Haffari, and Shirui Pan. 2024. Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. *Preprint*, arXiv:2410.13080.

A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. 2006. L-diversity: privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, Melbourne, Australia. Association for Computational Linguistics.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jinyoung Park, Minseok Joo, Joo-Kyung Kim, and Hyunwoo J. Kim. 2024. Generative subgraph retrieval for knowledge graph–grounded dialog generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21167–21182, Miami, Florida, USA. Association for Computational Linguistics.

Stefan Pasch and Min Chul Cha. 2025. Balancing privacy and utility in personal LLM writing tasks: An automated pipeline for evaluating anonymizations. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 32–41, Albuquerque, New Mexico. Association for Computational Linguistics.

Qwen. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Md Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. DialoKG: Knowledge-structure aware task-oriented dialogue generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2557–2571, Seattle, United States. Association for Computational Linguistics.

Mikhail Salnikov, Hai Le, Prateek Rajput, Irina Nikishina, Pavel Braslavski, Valentin Malykh, and Alexander Panchenko. 2023. Large language models meet knowledge graphs to answer factoid questions. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 635–644, Hong Kong, China. Association for Computational Linguistics.

Priyanka Sen, Sandeep Mavadia, and Amir Saffari. 2023. Knowledge graph-augmented language models for complex question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 1–8, Toronto, Canada. Association for Computational Linguistics.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2025. Large language models are advanced anonymizers. *Preprint*, arXiv:2402.13846.

Latanya Sweeney. 2002. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1855–1865, Hong Kong, China. Association for Computational Linguistics.

Derong Xu, Xinhang Li, Ziheng Zhang, Zhenxi Lin, Zhihong Zhu, Zhi Zheng, Xian Wu, Xiangyu Zhao, Tong Xu, and Enhong Chen. 2025. Harnessing large language models for knowledge graph question answering via adaptive multi-aspect retrieval-augmentation. *Preprint*, arXiv:2412.18537.

Zukang Yang, Zixuan Zhu, and Jennifer Zhu. 2025. CuriousLLM: Elevating multi-document question answering with LLM-enhanced knowledge graph reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 274–286, Albuquerque, New Mexico. Association for Computational Linguistics.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Yong Liu, Wei Chen, and Xiaoyan Zhu. 2021. EARL: Informative knowledge-grounded conversation generation with entity-agnostic representation learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2395, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4623–4629. International Joint Conferences on Artificial Intelligence Organization.

A Prompts

A.1 KG-DG Prompts

Prompts for dialogue generation for DeepSeek is in Tables 8 and 9. We utilized the same prompt for Qwen, removing <*think*> token.

```
You are an expert dialogue agent. Use the provided
conversation history and external knowledge (as
triplets) to generate a precise, fact-based reply.
The response should not be longer than 2-3 sentences
. Remember to stick to the knowledge.

History:
{history}

Knowledge:
{external_kg}

<think>
```

Table 8: Prompt for simple dialogue generation for LLMs.

```
You are a concise dialogue agent. Your task is to
generate a short, precise, and fact-based response
grounded strictly in the provided external knowledge
, which is given in the form of structured triplets.
 Use the conversation history to understand the
context, but only use facts that are directly
supported by the external knowledge when forming
your reply. Do not infer any information not
explicitly present in the external knowledge.

Guidelines:

- Limit your response to 2-3 sentences.
- Use clear and factual language.
- Do not invent or speculate beyond the given
knowledge.
- You may rephrase or combine knowledge triplets for
 naturalness, but do not introduce new facts.

Conversation History:
{history}

External Knowledge (Triplets):
{external_kg}

<think>
```

Table 9: A more detailed prompt for dialogue generation, emphasizing concise and fact-based responses grounded in external knowledge.

A.2 Anonymization Prompt

Our in-context learning prompt for anonymization is provided in Table 15.

A.3 Qualitative Analysis Prompt

Our qualitative analysis prompt is provided in Table 16.

A.4 LLM-KAT Entity Extraction

```
Your task is to extract, similar to the SQuAD V2
dataset, from the provided context that fills
triplets. The triplets are of format (Subject,
predicate, Object) in which either Subject or Object
 is missing. I use X instead of missing Subject or
Object in the triplet, you need to extract the
appropriate values for the X from the context. Each
extracted answer must be an exact substring from the
 context. If there are multiple valid answers,
separate them with @@. If no answer exists in the
context, output IS_IMPOSSIBLE.


Input Structure: The input is given in the following
 format: ID|||context|||triplet
There are 20 instances that you should response to.

Output Format: The final output should follow this
format: ID|||context|||triplet|||
answer_1@@answer_2@@answer_3 Replace
answer_1@@answer_2@@answer_3 with the actual answers
 found, or IS_IMPOSSIBLE if none are found. You
should output 20 lines.

Input:
{samples}

<think>
```

Table 10: Prompt for LLM-KAT entity extraction

Our prompt for LLM-KAT is provided in Table 10.

Task	GPU-Type	#GPUs	Time (H)
Dialogue Generation	I40s	4	2.322
LLM-KAT	I40s	4	3.759
UniEval	I40s	1	0.8
Anonymization	I40s	4	12.021

Table 11: Wall-on-clock time for experiments.

B Half Anonymized Dataset

We merge two anonymized and normal dataset by sampling 50% from each dataset. By doing so, we have a half anonymized dataset. The results for reasoning LLMs on this dataset are reported in Table 14.

Hyperparameter	DeepSeek	Qwen
Max new Tokens	16384	256
Top K	10	
Top P	0.95	
Temperature	0.6	
Batch Size	128	

Table 12: VLLM hyperparameters for Dialogue Generation.

Hyperparameter	Value
Seed	42
Max new Tokens	16384
Top K	10
Top P	0.95
Temperature	0.6
Batch Size	128

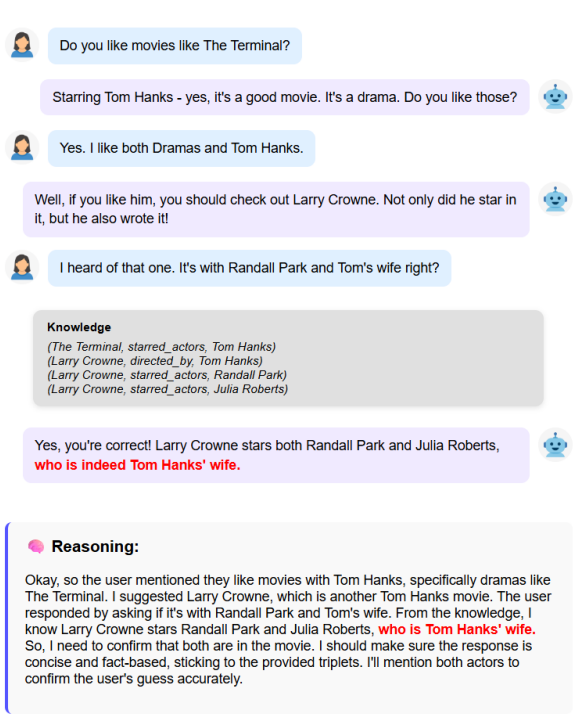
Table 13: VLLM hyperparameters for LLM-KAT.

C Case Study

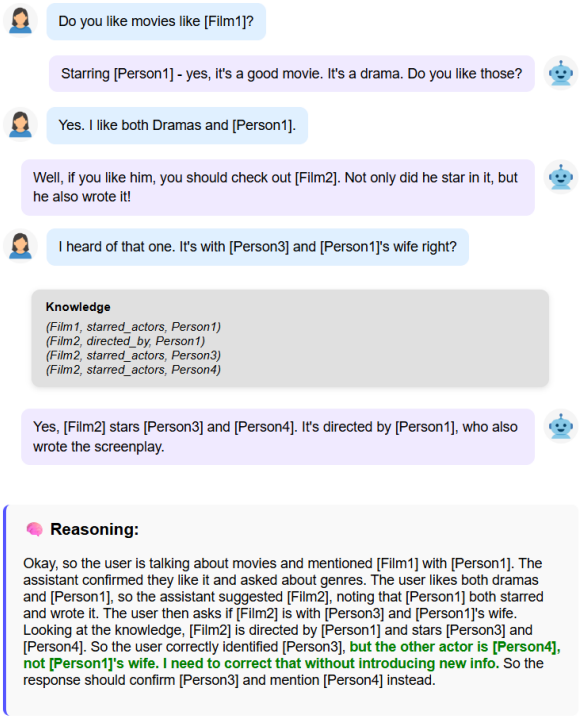
As shown in Figure 2a, the model made an assumption that Julia Roberts is the spouse of Tom Hanks. On the other hand, Figure 2b shows a correction in reasoning while processing Tom's wife and does not detach from the given knowledge. This case shows the effect of anonymization on the response generation that it prevents the model from misinformation generation.

Model	Dataset	LLM-KAT		UniEval	
		F1 Per Turn	F1 Per Session	Naturalness	Coherence
DeepSeek-r1-7B	Normal	76.04	77.16	92.36	98.59
	Half Anonymized	75.31	76.53	90.73	98.11
	Anonymized	76.37	77.77	88.89	97.75
DeepSeek-r1-14B	Normal	84.84	86.24	93.71	98.37
	Half Anonymized	86.15	87.61	92.66	98.25
	Anonymized	89.19	90.95	91.50	98.09
DeepSeek-r1-32B	Normal	85.56	86.80	94.83	99.02
	Half Anonymized	87.24	88.65	93.90	98.92
	Anonymized	89.75	91.46	92.90	98.84

Table 14: DeepSeek performance under normal, half-anonymized, and fully anonymized input on LLM-KAT and UniEval.



(a) Normal dataset input.



(b) Anonymized dataset input.

Figure 2: Case studies comparing (a) the normal dataset input and (b) the anonymized dataset input.

```
You are provided with a dialogue containing references to teams, people, locations, and literary influences,
along with a set of related knowledge triplets. Your task is to anonymize the dialogue and the knowledge by
replacing all real-world names with anonymized entity identifiers that include both a type (e.g., Team, Person,
Place, LiteraryInfluence) and an ID. Be sure to create a clear mapping table, update all occurrences in the
dialogue, and replace entries in the knowledge triplets accordingly.

Instructions:
1. Mapping Table:
   - First, scan the dialogue and the knowledge triplets to identify all unique entities (e.g., team names,
person names, places, literary influences).
   - For each entity, define:
     - An anonymized ID. For example:
       - Films: F1, F2, ...
       - Teams: T1, T2, ...
       - People: P1, P2, ...
       - Places: L1, L2, ...
       - Literary Influences or Authors: LI1, LI2, ...
     - Its type (Team, Person, Place, LiteraryInfluence, Film, etc).
   - Create a mapping table that shows each entity's original name, its anonymized ID, and its type.

2. Anonymize the Dialogue:
   - Replace every occurrence of a real-world entity in the dialogue with its corresponding anonymized ID.
   - Ensure that informal or abbreviated references (if any) are also mapped correctly.

3. Anonymize the Knowledge Triplets:
   - For each knowledge triplet, replace each entity with its corresponding anonymized ID based on the mapping
table.

4. Output Format:
   - Mapping Table: Present as a clear table with columns for ID, Original Entity, and Type.
   - Anonymized Dialogue: Provide the full dialogue with entity mentions replaced by their anonymized IDs.
   - Anonymized Knowledge Triplets: List the anonymized triplets exactly as in the original input, but with IDs
in place of the original entity names.

Example In-Context (from a previous task):

Mapping Table Example:
| **ID** | **Original Entity**                  | **Type**  |
|--------|--------------------------------------|-----------|
| F1     | Iron Man                             | Film      |
| P1     | Robert Downey Jr.                    | Person    |
| F2     | Zodiac (Crime Fiction Film)          | Film      |
| P2     | Jake Gyllenhaal                      | Person    |
| F3     | End of Watch                         | Film      |
| P3     | David Ayer                           | Person    |
| G1     | Thriller                             | Genre     |
| G2     | Crime Fiction                        | Genre     |

Anonymized Dialogue Example:
- S1: Do you like [F1]?
- S2: Sure do! [P1] is a favorite.
- S1: Yes, I like him too. Did you know he also was in [F2], a [G2] film?
- S2: I like [G2]! Didn't know [P1] was in there. [P2] starred as well.
- S1: So he did? He also starred in [F3]. Have you ever seen that movie?
- S2: Yes, I have! I like films directed by [P3]. How about you?
- S1: I have not. What genre is [F3]?
- S2: It's a [G1] and [G2] film as well.
- S1: I will make sure to check it out. I like [G1] films. Thank you!
- S2: Welcome!

Anonymized Knowledge Triplets Example:
1. ["[F1]", "starred_actors", "[P1]"]
2. ["[F2]", "starred_actors", "[P1]"]
3. ["[F2]", "starred_actors", "[P2]"]
4. ["[F3]", "~starred_actors", "[P2]"]
5. ["[F3]", "written_by", "[P3]"]
6. ["[F3]", "has_genre", "[G1]"]

Now, please perform the anonymization for the following query:

------------------------------
Dialogue:
{history}

Knowledge:
{external_kg}
------------------------------

Ensure that your final output includes:
- A Mapping Table with all identified entities, their anonymized IDs, and their types.
- The complete dialogue with anonymized entities.
- The knowledge triplets with all entities replaced by their anonymized IDs.

<think>
```

Table 15: Anonymization Prompt

```
You are a high-precision quality assessment agent. Your task is to evaluate two candidate responses (Option A and
 Option B) and determine which one is the superior choice--or whether both are equally valid--based on the
conversation history and structured external knowledge provided as triplets.

Follow these steps carefully:
    1. Understand the Context: Read the full conversation history to understand the dialogue intent, user queries
, and tone.
    2. Analyze the Knowledge: Examine the knowledge triplets, each in the format (subject, predicate, object).
Use them to fact-check and assess alignment with each option.
    3. Evaluate Each Option:
        - Check for factual correctness based on the triplets.
        - Make sure of the point that the response does not utilize any additional information/assumptions that
is not provided in the context.
    4. Make a Decision:
        - Choose A if Option A is better in using knowledge triplets.
        - Choose B if Option B is better in using knowledge triplets.
        - Choose Both if both are equally valid in using knowledge triplets.

Only output one of the following as your final decision: A, B, or Both. Just provide the final decision.

Conversation History:
{history}

Knowledge Triplets:
{knowledge_triplets}

Candidate Responses:
A: {A}

B: {B}
<think>
```

Table 16: Qualitative Analysis Prompt