Grounding LLMs to In-prompt Instructions: Reducing Hallucinations Caused by Static Pre-training Knowledge

Angus Addlesee
Heriot-Watt University
Edinburgh, UK
a.addlesee@hw.ac.uk

Abstract

When deploying LLMs in certain commercial or research settings, domain specific knowledge must be explicitly provided within the prompt. This in-prompt knowledge can conflict with an LLM's static world knowledge learned at pre-training, causing model hallucination (see examples in Table 1). In safety-critical settings, like healthcare and finance, these hallucinations can harm vulnerable users. We have curated a QA corpus containing information that LLMs could not have seen at pre-training. Using our corpus, we have probed various LLMs, manipulating both the prompt and the knowledge representation. We have found that our 'Jodie' prompt consistently improves the model's textual grounding to the given knowledge, and in-turn the overall answer accuracy. This is true in both the healthcare and finance domains – improving accuracy by up to 28% (mean: 12%). We have also identified that hierarchical and direct node-property graph structures could lead to more interpretable and controllable systems that provide a natural language interface with real-time in-domain knowledge. Our corpus will enable further work on this critical challenge.

Keywords: question answering, conversational AI, knowledge grounding, LLM evaluation, corpus

1. Introduction

LLMs are typically evaluated on their world knowledge learned at pre-training. For example, the popular Hugging Face Open LLM benchmark (the de facto standard leaderboard) ranks each model based on their performance across four tasks: (1) The AI2 Reasoning Challenge (Clark et al., 2018), a set of grade-school science questions; (2) MMLU (Hendrycks et al., 2020), a set of elementary level questions covering mathematics, US history, computer science, law, and more ; (3) HelloSwag (Zellers et al., 2019), testing whether the model can select "what will happen next?" given a common sense scenario and some options; and (4) TruthfulQA (Lin et al., 2022), a set of 817 questions on various topics, like law and politics, crafted to induce hallucinations due to common false beliefs.

These corpora (and others: FELM (Chen et al., 2023), HELMA (Li et al., 2023b), HaluEval (Li et al., 2023a), etc...), highlight the field's effort to reduce model hallucination. It is vital to clarify that they focus on hallucination reduction of outputs generated from the LLM's *static world knowledge*.

LLMs like ChatGPT and Bard are regularly asked questions in this manner, with users expecting the model to be an oracle of world knowledge. However, in both research and industry, these models are asked domain-specific questions (Neeman et al., 2023). For example, in a museum setting, a user might ask: "Can you tell me about exhibit 2?". An LLM-based dialogue system would only be able to answer correctly if the answer was provided in the prompt. This system may even state exhibit-related myths as facts *because* of its world knowledge. We are therefore interested in knowledge grounding to the in-prompt knowledge.

In this paper, we present the 'Jodie' prompt (full definition in Section 3). Our prompt provides the in-prompt knowledge as a quote by "Jodie W. Jenkins". The term "according to" has been shown reduce model hallucination to its static world knowledge (Weller et al., 2023), so following the quote, we ask the model to "Answer according to Jodie W. Jenkins". We hypothesise that this modification will encourage the model to respond using the in-prompt knowledge, rather than the static world knowledge.

LLMs are being applied in healthcare and finance, where this issue is critical. Google and Microsoft are using generative AI to answer doctor's questions (Capoot, 2023a,b), the EU's SPRING project is deploying LLM-based dialogue systems in a hospital to answer patient's questions (Addlesee et al., 2023b, 2024), and both BloombergGPT (Wu et al., 2023) and FinGPT (Yang et al., 2023) are released as robo financial advisors to users. In Table 1, we have provided examples from our corpus (see Section 2) with answers generated in this short paper. These examples highlight the hallucinations (in red) that have been generated *because of* the LLMs world knowledge learned at pre-training, and not grounded to the information provided in the prompt.

Allowing LLMs to add medical advice to its response could lead to real user harm. In response to Question (A) in Table 1, Llama-2-13b-chat adds details to its response that are not given in our hospital prompt. It is inconsistent, noting that you can drive both 24 and 8 hours after taking aspirin. *No* information about this time was provided in the

The Third Workshop on Safety for Conversational AI @LREC-COLING 2024, pages 1–7

Table 1: Two questions from our corpus. Some model responses are provided with hallucinations in red.

Question (A): Relevant Ground Text:		**Will I be ok to drive home?** As you will be given an aspirin, you cannot drive due to drowsiness.
Llama-2-13b-chat	Basic Prompt	You will not be able to drive home after your appointment as you will be given an aspirin, which can cause drowsiness. It is recommended that you do not drive for at least 24 hours after taking aspirin.
Llama-2-13b-chat	Expert Prompt	It is recommended that you do not drive for at least 8 hours after taking aspirin.
Llama-2-13b-chat	Jodie Prompt	You will be given an aspirin, which may cause drowsiness. Therefore, it is not advisable to drive home after the procedure.
Question (B): Relevant Ground Text:		**Is Twitter now called X?** Elon has been under heavy fire for his mismanagement of Twitter. The controversial renaming of the company to X has been a firestorm.
Llama-2-7b-chat	Basic Prompt	Twitter remains known as Twitter and has not been renamed to X.
Llama-2-7b-chat	Jodie Prompt	Twitter was recently renamed to X by its CEO, Elon Musk.

prompt, nor was it needed to answer the question. If a patient was given an incorrect duration, they may drive whilst still affected by the medication and have an accident. Our 'Jodie' prompt grounds to the given text in the prompt.

LLMs world knowledge is static. Therefore, even when given up-to-date info in the prompt, LLMs still hallucinate from their world knowledge. Llama-2-7b-chat consistently stated that Twitter's name has not changed, when asked Question (B) in Table 1, unless it was given our 'Jodie' prompt.

We have highlighted this prompt-grounding problem, and emphasised its safety-critical importance. We tackle it in this short paper using two methods: (1) Prompt engineering, manipulating the prompt; and (2) Knowledge engineering, manipulating the knowledge representation. We create a corpus and improve LLM answer accuracy by up to 28% in the healthcare setting, and 24% given financial reports.

2. Dataset Curation

As shown in Table 1, an LLMs world knowledge can conflict with domain specific prompt knowledge that can evolve in real-time. In order to evaluate LLM prompt grounding techniques, we need to provide information that was not seen by any LLM at pre-training. An LLM's exact pre-training data is often not public knowledge (Liesenfeld et al., 2023; Balloccu et al., 2024), so we curated two textual knowledge passages paired with 50 questions each (one in the healthcare domain, and one financial report). These were constructed in reverse order to each other, in case one method induced some unforeseen bias. Firstly, for the healthcare setting, we collated questions that real hospital patients asked a robot in a hospital memory clinic (Addlesee et al., 2023a,b). This SPRING corpus contains multi-party interactions between patients, their companions, and a social robot. Although this data was not released for question answering (QA), the captured interactions include many questions about directions, the cafe menu, hospital visiting hours, etc... The correct answers to these questions were not provided, and they would reflect a real hospital which an LLM may be familiar with (e.g. from its website). We therefore crafted a text passage that answers the 50 hospital related questions.

We created our finance QA data in the reverse order. We collated passages from three financial analysis documents from Seeking Alpha[1]. These were behind a paywall, and all the LLM answers were generated within 10 days of their publication. There is therefore no chance that the LLMs were pre-trained on these documents. The 50 questions were then human-generated from these texts.

The passage-question datasets were both the same in terms of passage length (600 words) and number of questions (50). Additionally, 70% of the questions in each domain are machine comprehension style, so the answer is a direct span of the given passage (e.g. "What is being served for lunch today?"). The other 30% require some additional reasoning (e.g. "How long until my appointment?", given the current time and appointment time in the passage). The main differences between the two domains is that the hospital data has a reading level of 7-8th grade (using the Dale-Chall readability formula, (Dale and Chall, 1948)), and contains very few named entities. Our finance data contains many people, stock ticker symbols, prices, and companies, which may induce more knowledge conflicts. Also, by the nature of financial analysis documents, the reading level was more complex, at graduate level (Dale and Chall, 1948).

In addition to prompt-engineering, we were keen to explore whether we can modify the knowledge representation itself to improve LLM prompt-grounding. We have therefore meticulously transformed the hospital passage information into a knowledge graph (KG) manually. A subset of this graph can be seen in Figure 1, visualised using GraphDB[2]. While LLMs are brilliant at language understanding and holding a wealth of general knowl-

[1] https://seekingalpha.com/

[2] https://graphdb.ontotext.com/

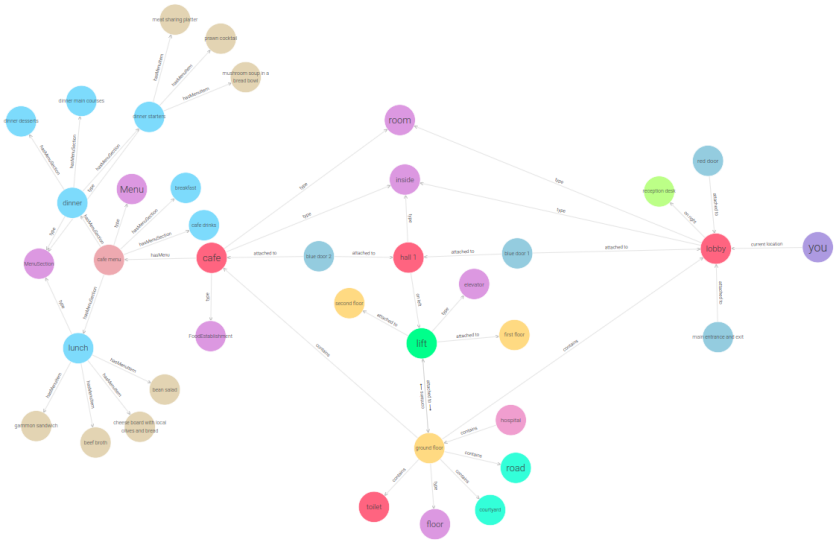

Figure 1: A subset of the hospital data represented as an RDF knowledge graph using Schema.org

edge, they hallucinate and lack domain specific or new knowledge. KGs, conversely, cannot understand natural language or unseen facts, but are excellent at providing an interpretable structure of domain specific knowledge that can evolve in real-time. If unified, these two technologies could be powerful (Pan et al., 2023; Ji et al., 2023).

Founded by Google, Microsoft, Yahoo!, and Yandex, Schema.org is the underlying structure of the internet. It is represented by the resource description framework (RDF, (Lassila et al., 1998; Manola et al., 2004)), which is used to describe KGs in triple statements. These technologies form the basis of Google, Wikipedia, Amazon Alexa, Facebook, eBay, and the list goes on... LLMs will have seen this data representation at pre-training. We have therefore created our hospital knowledge graph in RDF, using schema.org's ontology. We used a hierarchical data structure for the hospital cafe's menu, a multi-hop structure for directions, and relied on properties for reception and doctor information.

Our final corpus[3] contains a healthcare domain passage and knowledge graph paired with 50 questions that can be answered by either the text passage or KG directly. Additionally, the corpus contains a second text passage paired with 50 questions in the finance domain. Using this corpus, we can run prompt-grounding experiments via prompt-engineering in both domains, and knowledge-engineering in the healthcare domain.

3. Methodology

In related work, Weller et al. (2023) wanted to measure LLM's grounding to world knowledge. In this case, they selected Wikipedia as all LLMs will have seen this at pre-training. In order to measure how well an LLMs output grounded to Wikipedia, Weller et al. (2023) devised a metric: QUIP-score. This score is the character n-gram precision of the generated output compared to the source corpus. It is a useful metric in our case too, as we can measure how precisely each LLM's output is grounded in the given in-prompt knowledge. This focus on precision also punishes a model's output when it hallucinates, our goal of this paper. Using our corpus, we will use this QUIP-score and the answer's accuracy to measure prompt-grounding performance. Grounding is impractical if it does not preserve QA performance.

As LLMs are pre-trained on many news articles, the phrase "according to" has been shown to improve world knowledge grounding (Weller et al., 2023). Our 'Jodie' prompt is designed as a modification of this approach – instead aiming to improve in-prompt knowledge grounding by asking the model to answer according to a quote by "Jodie W. Jenkins". We provide four prompts:
Basic: The passage followed by the question.
Jodie: Our prompt provides the passage as a quote by Jodie W. Jenkins, a fictitious non-celebrity name (according to Google). We then ask the LLM to answer according to Jodie. The exact pattern is this: 'Jodie W. Jenkins said "PASSAGE". Answer according to Jodie W. Jenkins. QUESTION'.
Expert: In order to ensure any prompt-grounding benefit is not simply a result of adding "according to", we again provide the passage as a quote by Jodie W. Jenkins, but add "Answer according to Bloomberg" instead of Jodie in the finance domain ("UnitedHealth" in the healthcare domain).
Wikipedia: The Expert prompt with one word replaced. The expert name is set to "Wikipedia".

[3] https://github.com/AddleseeHQ/in-prompt-grounding

Table 2: Healthcare results. ■ indicates an improvement compared to the 'basic' prompt. ■ indicates a performance drop compared to the 'basic' prompt. **Bold** marks the best scores per model.

LLM	Basic Prompt		Jodie Prompt		Expert Prompt		Wikipedia Prompt	
	Quip	Acc	Quip	Acc	Quip	Acc	Quip	Acc
Dolly-12b	38.71	36	35.74	**42**	28.08	32	**39.21**	34
GPT-4	41.04	94	**42.92**	**98**	42.61	92	38.66	90
Llama-7b-chat	43.06	56	**44.56**	**84**	41.64	72	40.84	74
Llama-13b-chat	**48.51**	**60**	41.18	**60**	44.04	50	44.29	58
Llama-70b-chat	44.10	64	**58.73**	**82**	52.44	70	53.78	68
Llama-70b-chat (0.95 temp)	44.52	68	**53.18**	**80**	52.01	70	52.82	68
Vicuna-13b-v1.1	64.93	46	**80.95**	**54**	29.17	12	31.93	26
Vicuna-13b-v1.5	40.97	70	**41.14**	**74**	36.30	52	34.17	56

Table 3: Finance results with the same visual key as Table 2.

LLM	Basic Prompt		Jodie Prompt		Expert Prompt		Wikipedia Prompt	
	Quip	Acc	Quip	Acc	Quip	Acc	Quip	Acc
Dolly-12b	14.07	20	**20.24**	**30**	19.19	18	13.82	24
GPT-4	**37.39**	74	36.55	**82**	36.08	74	31.04	68
Llama-7b-chat	40.91	68	**46.15**	**76**	42.69	62	37.96	62
Llama-13b-chat	42.95	68	**43.10**	**74**	37.67	62	40.17	64
Llama-70b-chat	45.41	64	**52.76**	**80**	49.88	70	45.05	62
Llama-70b-chat (0.95 temp)	45.38	62	**54.36**	**82**	47.97	68	47.31	58
Vicuna-13b-v1.1	43.65	44	**61.33**	**64**	39.53	34	22.55	30
Vicuna-13b-v1.5	32.55	46	**56.08**	**70**	53.52	62	47.24	48

4. Results

Using our new corpus, we evaluated various LLMs hosted by Replicate, through their API (excluding GPT-4, for which we used OpenAI's API) with the metrics and prompts described in Section 3. The LLMs evaluated were: Dolly-12b, GPT-4, Llama-2-7b-chat, Llama-2-13b-chat, Llama-2-70b-chat (Touvron et al., 2023), Vicuna-13b-v1.1, and Vicuna-13b-v1.5 (Chiang et al., 2023). We set each model temperature to 0.4 for more deterministic results, but additionally ran all the experiments with Llama-2-70b-chat's temperature set to 0.95.

Prompt Engineering:
In the healthcare domain, Table 2 illustrates the impressive performance of our 'Jodie' prompt. The Quip-score did decrease for two of the models, but the accuracy never deteriorated, and increased by up to 28% (mean: 10%). Even though the 'Expert' and 'Wikipedia' prompts differ from the 'Jodie' prompt by just one name, they generate more text that is not contained in the given prompt (as shown by the lower Quip-scores), and these additional hallucinations result in an accuracy drop. While this paper is not comparing the models to each other, GPT-4's performance is remarkable, particularly its accuracy in the healthcare domain.

In the finance domain, with a more complex text that contains numerous named entities, these findings are even more evident. Table 3 shows large boosts to both the Quip-score and answer accuracy when given our 'Jodie' prompt. The accuracy increased by up to 24% (mean: 14%), and the other prompt's poor performance shows that the boost is not due solely to the 'according to' phrase.

Knowledge Engineering:
As detailed in Section 2, integrating LLMs with knowledge graphs (KGs) will lead to more interpretable and controllable systems that enable a natural language interface with real-time in-domain knowledge. Commercial systems are being announced (e.g. Stardog Voicebox (Grove, 2023) or the OpenLink Virtual Assistant (Uyi Idehen, 2023)), but at time of writing, they are not publicly available.

Instead of providing the hospital information to each LLM as a text passage, we passed each LLM the KG in our corpus, and asked each of the healthcare questions. The entire KG was too big for most of the LLM's prompt size limits, so we split the KG into four subgraphs: the directions, the cafe info, the reception info, and the doctor info. The hospital questions were sourced from interactions with a modular dialogue system (Addlesee et al., 2023b) with similar question categories, like their 'directions' and 'reception' bots (Gunson et al., 2022).

Using our KG, we passed all 50 hospital questions to each LLM along with the relevant subgraph. GPT-4 has a larger prompt size, so we also evaluated it whilst providing the full KG with each question, indicated by '(full)' in the table. The basic prompt simply provided the KG and the question. The 'Grounding' prompt used the 'Jodie' prompt method again. The results are in Table 4, and we omit Dolly and Vicuna-13b-v1.1 due to their poor performance (full row of zeros), we do not recommend using them if your data is stored as a KG.

Once again, the grounding prompt improved overall performance. As information in the graph was structured differently, we report the results per

Table 4: Knowledge graph results using the hospital KG in our corpus. Reporting answer accuracy.

LLM	Total Acc (N=50)		Directions Acc (N=13)		Cafe Acc (N=13)		Reception Acc (N=13)		Doctor Acc (N=11)	
	Basic Prompt	Grounding Prompt	Basic Prompt	Grounding Prompt	Basic Prompt	Grounding Prompt	Basic Prompt	Grounding Prompt	Basic Prompt	Grounding Prompt
GPT-4 (full)	84	86	83.3	91.7	100	92.3	69.2	76.9	81.8	81.8
GPT-4	84	88	83.3	100	100	100	69.2	69.2	81.8	81.8
Llama-7b-chat	30	46	8.3	25.0	38.5	76.9	38.5	30.8	27.3	45.5
Llama-13b-chat	46	52	16.7	8.3	53.8	76.9	61.5	61.5	45.5	54.5
Llama-70b-chat	62	66	16.7	33.3	76.9	76.9	76.9	69.2	72.7	81.8
Vicuna-13b-v1.5	44	46	33.3	16.7	46.2	46.2	38.5	61.5	54.5	54.5

question type. The LLMs performed particularly well when asked cafe related questions. We modelled cafe knowledge using a hierarchical structure, which the LLMs have clearly learned to parse. To answer the direction questions accurately, the LLM had to follow multiple graph edges, hopping through nodes to find a path from one location to another. This structure was suboptimal, and the larger Llama models struggle with this in particular. The reception and doctor knowledge was modelled using many node and class properties, but there was a notable difference. The doctor information relied on node properties, which the LLMs parsed well. The reception knowledge relied on class properties, which even GPT-4 struggled with more. To clarify, we did not annotate every hospital location with the 'smokingAllowed' property. We ascribed each location to one of two classes: 'Inside' or 'Outside'. These classes were then connected to the smoking property. Therefore, when asked if it was allowed to smoke in the courtyard, the LLM had to reason that the courtyard is a member of the 'Outside' class, and smoking is therefore allowed. We recommend using the more repetitive node properties and a hierarchical structure. This could be done at the data modelling stage, or at runtime using an RDF reasoning engine, like RDFox (Nenov et al., 2015), on the intermediate representation.

5. Conclusions and Future Work

In this short paper, we highlight the safety-critical issue of LLM grounding to the in-prompt knowledge given at runtime. We show that when LLMs use their world knowledge learned at pre-training to answer a question, it can lead to hallucination due to the specific domain, or the world knowledge being out of date. We created a corpus of two text passages and a KG representing knowledge in the healthcare and finance domains. This information could not have been seen by any LLM, and 50 questions were paired with each domain.

Our 'Jodie' prompt consistently grounded LLM answers to the given in-prompt knowledge, and this increased accuracy up to 28% (mean: 12%). The same prompt-engineering method worked when given a KG in the prompt. The KG did result in lower accuracy scores overall, but we found that hierarchical and direct node-property edges were better structures to use with LLMs. We believe the integration of KGs and LLMs will ultimately lead to interpretable systems that enable a natural language interface with real-time in-domain knowledge.

Ethical Consideration

Knowledge grounding is critical for LLM safety, particularly in domains like healthcare and finance. We have presented methods that anyone could implement effortlessly today with other methods like guardrails and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). Fine-tuning provides another approach, but recent work suggests that this can inadvertently reduce the effectiveness of LLM safety guardrails (Qi et al., 2023). This poses a dilemma in sensitive domains.

Considering again the driving after aspirin example found in Table 1, we successfully poisoned the prompt to provide an incorrect answer of 3 hours. Through dialogue, a bad actor can manipulate the LLM to output a harmful response to a vulnerable user. This must be considered if deploying an LLM in the wild. Deleting dialogue history, or resetting the context between users, could mitigate this risk.

Finally, all of our questions were in-domain. That is, they could be answered given the prompt knowledge. Our work aimed to improve grounding to the in-prompt knowledge, so this was the scope of the short paper. We did try asking various out-of-domain questions given the 'Jodie' prompt. Trivia questions and joke requests were still answered, but in the hospital setting, questions like "What is my age?" and "Where is the radiology department?" were thankfully not answered (no information about radiology is provided in the prompt). This is promising, but we recommend further testing out-of-domain questions that are specific to your setting before deploying our prompt.

Acknowledgements

This research was funded by the EU H2020 program under grant agreement no. 871245 (https://spring-h2020.eu/). We would also like to thank Replicate and Seeking Alpha for supporting this work.

Bibliographical References

Angus Addlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernández García, Nancie Gunson, Weronika Sieińska, Christian Dondrup, and Oliver Lemon. 2024. Multi-party multimodal conversations between patients, their companions, and a social robot in a hospital memory clinic. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon. 2023a. Data collection for multi-party task-based dialogue in social robotics. In *Proceedings of the 13th International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon. 2023b. Multi-party goal tracking with llms: Comparing pre-training, fine-tuning, and prompt engineering. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv:2402.03927*.

Ashley Capoot. 2023a. Google announces new generative ai search capabilities for doctors. *CNBC*.

Ashley Capoot. 2023b. Microsoft announces new ai tools to help doctors deliver better care. *CNBC*.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: Benchmarking factuality evaluation of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Mike Grove. 2023. Llm will accelerate knowledge graph adoption. *Stardog*.

Nancie Gunson, Daniel Hernández García, Weronika Sieińska, Christian Dondrup, and Oliver Lemon. 2022. Developing a social conversational robot for the hospital waiting room. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1352–1357. IEEE.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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.

Ora Lassila, Ralph R Swick, et al. 1998. Resource description framework (rdf) model and syntax specification. *W3C recommendation*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv e-prints*, pages arXiv–2305.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Helma: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–6.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Frank Manola, Eric Miller, Brian McBride, et al. 2004. Rdf primer. *W3C recommendation*, 10(1-107):6.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070.

Yavor Nenov, Robert Piro, Boris Motik, Ian Horrocks, Zhe Wu, and Jay Banerjee. 2015. Rdfox: A highly-scalable rdf store. In *International Semantic Web Conference*, pages 3–20. Springer.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying large language models and knowledge graphs: A roadmap. *arXiv preprint arXiv:2306.08302*.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Kingsley Uyi Idehen. 2023. Introducing the openlink virtual assistant. *Openlink Software*.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2023. " according to..." prompting language models improves quoting from pretraining data. *arXiv preprint arXiv:2305.13252*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.