Enhancing Recommendation Systems Using Large Language Models and Personalized Knowledge Graphs

Fernando Spadea[1]

[1]Rensselaer Polytechnic Institute, Troy, New York, USA

Abstract

We investigate how large language models (LLMs), when paired with personalized knowledge graphs (PKGs), can power decentralized recommendation systems. This work lays the groundwork for more intelligent and user-aligned personal digital assistants that respect user autonomy and data sovereignty. A key focus of our research is exploring how LLMs can be fine-tuned in federated settings to balance personalization with privacy. To this end, we evaluate several fine-tuning methods and compare their performance to select the best one. Early results indicate that LLMs fine-tuned to use PKGs can outperform symbolic and embedding-based KGC models (e.g., KBGAT, HAKE) in both centralized and federated contexts, and that fine-tuning with Kahneman-Tversky Optimization (KTO) is more resilient to lopsided data distributions.

Keywords

Knowledge Graph, Federated Learning, Knowledge Graph Completion, Personalized Knowledge Graphs, Kahneman-Tversky Optimization, Fine-Tuning, Large Language Model

1. Problem Statement

Current recommendation systems are centralized, opaque, and dependent on collecting and analyzing large volumes of user data. This creates several critical challenges:

- **Privacy Risks**: Users must relinquish control over personal data to centralized servers, often owned by large companies. In addition to trusting these companies to use their data responsibly, they must also trust them to protect their data. These servers become high-value targets for malicious actors, increasing the risk of data breaches and unauthorized surveillance.
- **Scalability Constraints**: As the user base grows, centralized data infrastructures must scale accordingly, demanding increasing storage, computational power, and bandwidth. This makes it difficult for smaller companies or organizations with limited resources to develop or deploy recommendation systems.
- **Lack of Interpretability and Control**: Most current systems operate as black boxes, offering little transparency or control to users. As a result, users are often unaware of why certain content is recommended to them and have minimal ability to influence or adjust the underlying decision process. This can reinforce Personalized Information Environments (PIEs), where recommendations are narrowly tailored, limiting exposure to new content and creating a filter bubble.

We propose a privacy-preserving, scalable, and semantically interpretable recommendation system using a Large Language Model (LLM) trained via federated learning (FL) [1]. Instead of aggregating raw user data centrally, our approach enables local, privacy-aware model training. Each user maintains a personalized knowledge graph (PKG), a human- and machine-readable structure of entities and relations based on user interactions and preferences. The PKG evolves over time as the user engages with the system, particularly through conversations with the LLM.

For instance, consider a user interested in movie recommendations. Their PKG might initially contain a preference for ”Raiders of the Lost Ark (1981)”. After they tell their assistant they enjoy Harrison

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

spadef@rpi.edu (F. Spadea)
0009-0006-4278-3666 (F. Spadea)

Ford's work, the system recommends another of his movies, like "What Lies Beneath (2000)". If the user accepts this suggestion, their PKG is updated to reflect this new preference. Later, when the user's device participates in a round of federated learning, the model is fine-tuned locally on this new interaction. The resulting model updates are sent to a central server for aggregation. Because the user's specific preferences never leave their device, their privacy is preserved, while the aggregated global model still learns general patterns, such as the connection between different Harrison Ford movies, benefiting all users.

Because PKGs are interpretable and adaptable, users can both understand and influence how their data informs recommendations. The system can be flexibly tuned to specific recommendation contexts. They can also be designed to deliberately disrupt PIEs by adapting preference patterns within the PKG.

2. Importance

Why is this problem important and for whom?

This problem has broad societal relevance, impacting virtually all users of modern digital platforms. For individuals, it concerns the fundamental right to privacy, transparency, and control over personal data. Furthermore, exploring a variety of content enriches a user's experience. For organizations, especially smaller companies and startups, it presents a barrier to entry due to the infrastructure requirements and legal risks of handling sensitive user data.

Who will benefit and who should care?

- Users gain enhanced privacy, transparency, and agency over how they are profiled and marketed to.
- Companies and organizations benefit from reduced infrastructure costs and data liability, making advanced recommendation systems more accessible to smaller players.
- Researchers interested in privacy, personalization, fairness, and human-computer interaction gain a novel paradigm for user modeling and recommendation.

What is the impact of solving this problem (for the research community, or society in general)?

Solving this problem could democratize access to recommendation technology while restoring trust in digital systems. It fosters user empowerment through interpretable data structures and safeguards personal data by eliminating the need for centralized storage. From a societal perspective, it combats the dangers of algorithmic echo chambers by enabling users to actively shape their digital experiences. From a research perspective, it introduces a hybrid model that combines language-based reasoning with structured, user-owned knowledge, contributing to the evolution of federated and interpretable AI. This approach aligns with recent perspectives that unifying KGs and LLMs can combine their respective strengths: the factual knowledge and interpretability from KGs with the advanced reasoning and language abilities of LLMs [2].

3. Related Work

Has a solution to this problem been attempted before and how?

Prior work has explored key components of our proposed system—namely, knowledge graph completion (KGC), FL, and recommendation using LLMs, but no existing approach integrates these elements into a unified, interpretable, privacy-preserving, and scalable recommendation framework.

In the space of centralized KGC, models such as KBGAT [3] and HAKE [4] have achieved strong Hits@k performance by leveraging graph embeddings and attention mechanisms. Federated approaches to KGC are rare but do exist; notably, Fede [5] explores federated training of KGC models across distributed clients.

To test the potential of fine-tuning LLMs with FL, Ye et al. [6, 7] designed a set of benchmarks, named OpenFedLLM, to test the capabilities of federated LLMs. They tested several fine-tuning methods across several models and federated aggregation methods; most notably, they tested Direct Preference Optimization (DPO), a state-of-the-art human-feedback based fine-tuning method, and found that it performed well [8].

LLMs have also been tested for their ability to perform KGC tasks. For example, Meyer et al. [9] experimented with using ChatGPT for KGC, though in a limited, non-federated setting. In a complementary direction, Qiu et al. [10] incorporate knowledge graphs (KGs) to improve LLM-based recommendation, but without personalization or user-side control of data.

Efforts to mitigate PIEs include the Dual Echo Chamber framework [11], which models both the user's comfort zone and alternative spaces using KG embeddings. Anand et al. [12] propose calculating data point influence to selectively retrain models in order to diversify recommendations.

If you are addressing an existing problem, what are the limitations of current solutions?

Despite their strengths, existing KGC models like KBGAT [3] and HAKE [4] are unsuitable for our setting due to several key limitations. These models assume centralized training and storage, requiring access to the global list of entities and relations when the model is built. Not only does this violate the constraints of federated environments, but it also requires that the models be rebuilt and retrained whenever new entities or relations are introduced. This is especially problematic in a federated setting where clients may leave the network and thus take their data with them so that it can no longer be used in the retraining of the model. Additionally, current models cannot constrain individual KGC tasks to specific subdomains (e.g., food, media), limiting their usefulness in personalized, on-demand recommendation scenarios.

While OpenFedLLM [6, 7] does provide a useful framework and set of benchmarks for testing federated LLMs, it has a significant limitation: it predates Kahneman-Tversky Optimization (KTO) [13] and does not cover or support it as a result. KTO is a newer human-feedback fine-tuning method, like DPO, but even more widely applicable and effective, making it even more appropriate for FL, so this is a significant drawback. There is very limited testing of KTO in FL overall, so there is significant room for exploring its performance in a federated setting.

While Fede [5] explores KGC in a federated setting, it does so globally and without personalized structures like PKGs.

Regarding LLM-based solutions, existing work does not integrate federated training or PKGs. Most rely on static prompting rather than fine-tuning, and do not explore dynamic, user-guided recommendations or PIE mitigation strategies.

Efforts to explore outside PIEs also suffer limitations. The Dual Echo Chamber framework [11] requires case-specific strategies for different PIE types, which is not scalable. Additionally, the retraining-based method by Anand et al. [12] is computationally expensive and lacks real-time adaptability.

What are you adding that is novel? Why? If not, have research efforts tried or solved similar, analogous problems?

Our proposed system introduces several novel contributions:

- **Federated, PKG-based Recommendation**: We integrate federated learning with personalized KGs, allowing users to retain control over their data while still benefiting from high-quality, LLM-driven recommendations. Additionally, we employ differential privacy to ensure the user's data remains private [14].
- **LLM Fine-Tuning with KTO**: We leverage KTO to fine-tune the LLM for knowledge graph tasks, enabling more effective and context-aware recommendations [13]. Among fine-tuning variants of DPO, KTO is found to produce the best results [15].

- **Guideable and Contextual Recommendations**: Our system allows users to guide recommendations toward specific domains by instructing the LLM to work within said domains, improving relevance and interpretability.
- **Real-time PIE Mitigation**: Unlike existing approaches that require retraining or heuristic strategies, we address PIEs directly by modifying the user's PKG, enabling quick, targeted adjustments without retraining the model.

This combination of techniques results in a system that is privacy-preserving, interpretable, scalable, and user-controllable, qualities that are not achieved in existing work.

What can you learn from these efforts?

These prior efforts underscore the growing interest in privacy, personalization, and interpretability in recommendation systems. They highlight both the promise and limitations of existing technologies. Specifically, they show: (i) The potential of LLMs in semantic reasoning tasks involving KGs. (ii) The practical challenges of deploying centralized models at scale. (iii) The importance of user-guided and interpretable systems in exploring new content. Together, these works provide valuable benchmarks and conceptual foundations upon which we can build a more robust and user-centered solution.

4. Research Questions and Hypothesis

What research questions do you plan to explore?

We aim to investigate how LLMs can be integrated into a privacy-preserving, decentralized recommendation system grounded in user-controlled knowledge representations. Specifically, our work explores the following research questions:
RQ1: How can PKGs be used to fine-tune an LLM to improve recommendation accuracy and personalization compared to standard knowledge-aware recommenders? (Core technical question focusing on PKG+LLM efficacy)
RQ2: Which fine-tuning strategies are most effective for LLMs in a federated learning setting, particularly in the presence of heterogeneous data and privacy constraints?
RQ3: How does knowledge graph completion via LLM, with PKG data, impact recommendation quality, such as coverage of long-tail items and novelty?

These questions leverage Semantic Web technologies to structure and store PKGs in interoperable, machine-readable formats. By grounding user data in open standards (e.g., JSON-LD), stored within Solid pods, we aim to ensure that users retain full control over their data while still enabling high-performance model inference on the client side.

What hypotheses do you make in formulating your solution?

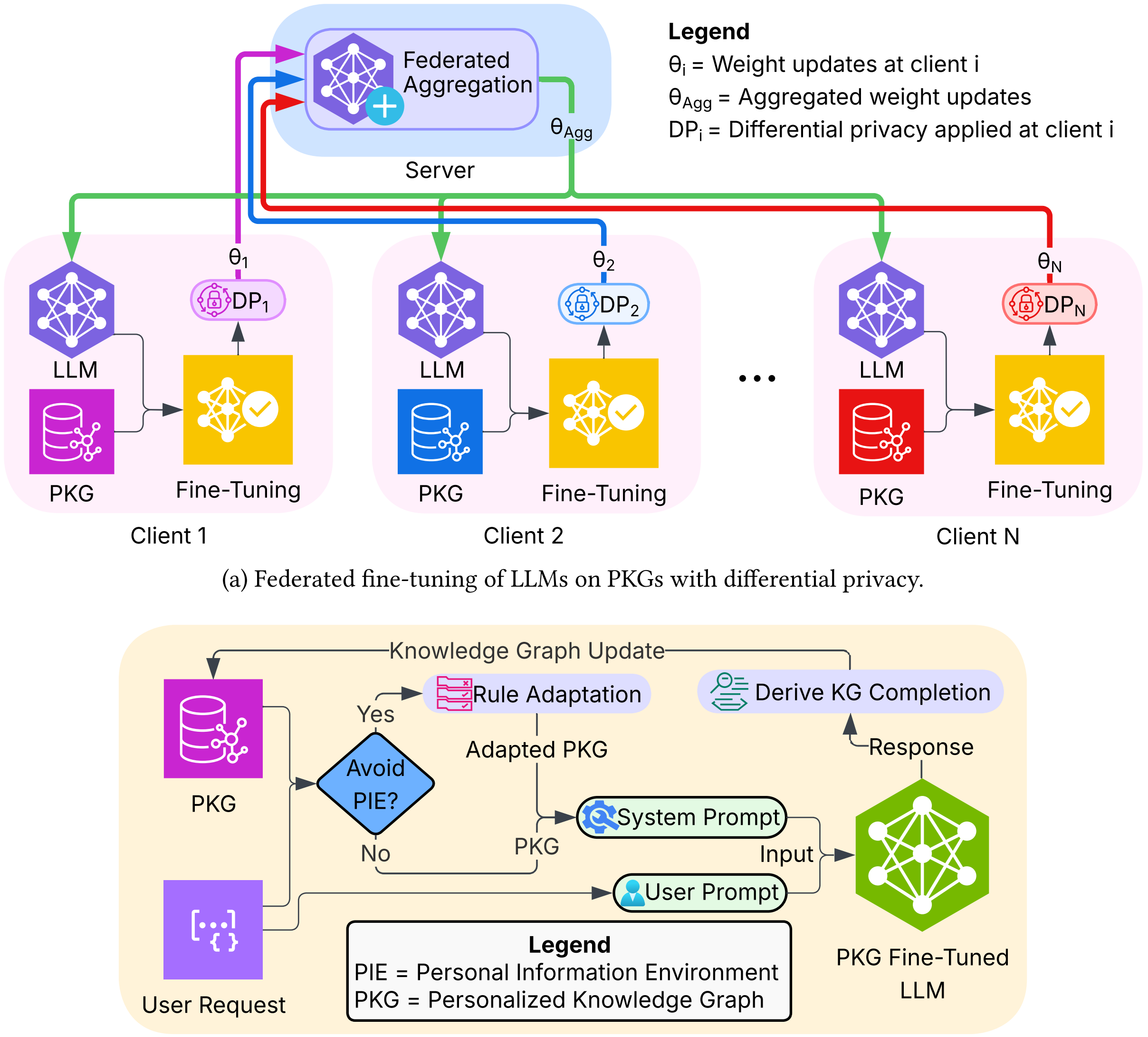
H1: PKGs will enable an LLM to generate more accurate recommendations than an LLM without KG fine-tuning, due to the injection of structured personal knowledge.
H2: LLMs fine-tuned with KTO are more performant and resilient to lopsided data in a federated setting, even with differential privacy applied, than those fine-tuned via other human feedback-based fine-tuning methods, such as Direct Preference Optimization (DPO) [8].
H3: If a user's PKG is adjusted, an LLM fine-tuned for KGC can be led to make recommendations outside the PIE without retraining more effectively than if the LLM was simply asked to avoid the PIE with the unadjusted PKG.

5. Preliminary Results

Do you have any preliminary results that inform your research questions or hypotheses?

Figure 1 illustrates our framework design. Figure 1a shows how clients perform local training on personalized KGs and communicate model updates to a central server for aggregation. Before sending the updates, differential privacy is applied to ensure the user's data cannot be reverse engineered from the updates. Figure 1b shows how, at inference time, the LLM receives a user's PKG and natural language query, and generates KG completions. However, if the user indicates they want to avoid the PIE in the request, the PKG is first adjusted to represent a user with PIE-avoiding preferences.

Table 1 contains our previous work [16] where we showed that an LLM fine-tuned, via FL, with KTO outperformed DPO, even when the distribution of the data was randomized. We fine-tuned alpha7B [17] with chatbot arena data [18] in a federated setting using both DPO and KTO, with and without randomizing the data (KTOR and KTOO respectively). Our findings demonstrate that KTO consistently outperforms DPO across all aggregation methods and evaluation benchmarks. Notably, KTO achieves higher scores even when the training data is randomized (KTOR), highlighting its robustness to data distribution variance. The evaluated benchmarks include:



(a) Federated fine-tuning of LLMs on PKGs with differential privacy.

(b) PIE-aware recommendation with rule adaptations.

Figure 1: Overview of the proposed PKG completion framework.

Table 1
Benchmark Results: KTOR and KTOO represent models trained with KTO with and without data redistribution, respectively. MMLU and AdvBench are scored out of 100, others out of 10. Green arrows indicate which fine-tuning method performed best.

Aggr. Method	MT-Bench-1 (/10)			Vicuna (/10)			AdvBench (/100)			MMLU (/100)		
Base	7.51			7.51			9.62			25.70		
	DPO	**KTOO**	**KTOR**	**DPO**	**KTOO**	**KTOR**	**DPO**	**KTOO**	**KTOR**	**DPO**	**KTOO**	**KTOR**
FedAvg	7.84	8.14 ↑	8.11	8.03	8.51 ↑	8.40	12.50	15.77 ↑	12.69	28.56	32.35 ↑	32.13
FedProx	7.73	8.44 ↑	8.01	7.73	8.39 ↑	8.23	13.08	16.15 ↑	15.58	28.32	32.47	34.34 ↑
SCAFFOLD	8.01	8.17 ↑	7.83	7.91	8.34 ↑	8.21	14.23	14.62	17.50 ↑	30.11	31.71 ↑	31.29
FedAvgM	7.16	7.54	7.56 ↑	7.84	7.99	8.37 ↑	8.65	12.31	14.81 ↑	27.35	25.55	33.58 ↑
FedYogi	8.75	8.98	9.03 ↑	7.65	8.21 ↑	8.13	11.35	12.88	17.12 ↑	26.75	28.24 ↑	28.15
FedAdagrad	8.49	8.84 ↑	8.78	7.96	8.32	8.34 ↑	11.54	12.88 ↑	11.92	26.94	27.85 ↑	27.57
FedAdam	8.20	8.64 ↑	8.43	7.89	8.55 ↑	8.47	11.35	12.69	13.46 ↑	26.48	27.98 ↑	27.95

Table 2
Model performance on the Movie KG dataset.

Model	Precision	Recall	Hits@1	Hits@3	Hits@10	MRR
PKGLLM (NoSyn)	0.4920	0.2012	0.2342	0.2384	0.2395	0.2364
PKGLLM	**0.5486**	0.2110	0.2492	0.2506	0.2512	0.2500
FedPKGLLM (NoSyn)	0.4603	**0.4088**	**0.4710**	**0.4835**	**0.4838**	**0.4771**
FedPKGLLM	0.3488	0.3924	0.4116	0.4604	0.4641	0.4361
KBGAT	0.0277	0.2288	0.1495	0.1816	0.2288	0.1797
KBGAT-fed	0.0017	0.0165	0.0019	0.0072	0.0165	0.0084
HAKE	0.0084	0.0802	0.0130	0.0324	0.0802	0.0380
HAKE-fed	0.0022	0.0222	0.0000	0.0055	0.0222	0.0087

- **MT-Bench-1**: Measures one-turn conversational performance [18].
- **Vicuna**: Assesses instruction-following capabilities [19].
- **AdvBench**: Evaluates model safety and adversarial robustness [20].
- **MMLU**: Measures factual knowledge and reasoning ability [21].

The MT-Bench-1, Vicuna, and AdvBench results have been published in our prior work [16], while the MMLU results are newer and not yet published.

Table 2 reports performance on our custom Movie KG dataset, derived from the *Recommendation Dialogues* dataset [22, 23]. We took dialogues between users recommending movies to build PKGs that encode each user's preferences. In the table, we refer to our model as PKGLLM (centralized) and FedPKGLLM (federated) with (noSyn) indicating that no synthetic data was used in fine-tuning.

We evaluated LLMs fine-tuned with KTO against strong symbolic and embedding-based KGC baselines: KBGAT and HAKE. Results show that our model, PKGLLM, which is a KTO-tuned Qwen3-0.6B model [24], achieves significantly better performance in both centralized and federated configurations, across all metrics: Precision, Recall, Hits@K, and MRR. This suggests that LLMs trained with KTO can outperform specialized KGC models on link prediction tasks grounded in personalized, dialogue-derived KGs. These results strongly support **H1** and **H2**, indicating that LLMs fine-tuned via KTO can serve as effective and robust KGC engines, particularly in decentralized, privacy-sensitive settings.

6. Evaluation

How will you know you've answered your question(s)?

Our evaluation strategy centers on systematically testing each hypothesis and research question through comparative experiments and controlled ablation studies. For RQ1 and RQ2, we rely on benchmark datasets and model performance metrics as outlined in the preliminary results. These results already support **H1** and part of **H2**, showing that LLMs fine-tuned via KTO outperform DPO-based models in a federated setting across a variety of benchmarks, and an LLM fine-tuned to make recommendations using PKGs outperforms KBGAT and HAKE. For further testing, we will continue to compare against baseline models.

What are the methods you apply to test your hypotheses?

To evaluate the other half of **H2** (regarding the effect of differential privacy), we will fine-tune Alpaca7B using both KTO and DPO under varying levels of privacy budgets. We select Alpaca7B to maintain comparability with the OpenFedLLM [6, 7], and also because more advanced models may actually decrease in performance from fine-tuning on the chatbot arena dataset since they likely have already used it in training. Performance will be assessed using similar metrics as in Table 1, allowing for a direct comparison of robustness under privacy constraints.

To assess **H3** and investigate the impact of KG adaptation on recommendations that venture outside a user's PIE, we will conduct experiments using three strategies:

1. Recommendations generated with the unmodified PKG.
2. Recommendations generated by prompting the model to avoid the PIE without PKG adaptation.
3. Recommendations generated after adapting the PKG to avoid the PIE.

We will compare the precision, relevance, and diversity of recommendations with each strategy to determine the efficacy of PKG adaptation as a strategy for PIE avoidance.

Have you identified criteria to measure the degree of success of your solution?

For hypotheses related to differential privacy (**H2**), we will use the same benchmarks as in the preliminary results: MT-Bench-1, Vicuna, AdvBench, and MMLU. These metrics provide comprehensive coverage of conversational quality, instruction following, safety, and factual knowledge.

For evaluating PIE avoidance (**H3**), we will categorize recommendations into three buckets:

- **Invalid Recommendation**: The recommendation fails to satisfy the user's explicit query constraints
- **In-PIE Recommendation**: The recommendation is valid but reinforces the user's existing PIE, offering little novelty.
- **Out-PIE Recommendation**: The recommendation is valid and successfully steers the user away from their established PIE.

Consider two illustrative cases: (1) a user seeking a cute animal video but aiming to avoid cat videos, having been overexposed to them; and (2) a user searching for Italian food recipes while wanting to avoid tomatoes due to over-personalized exposure. Table 3 presents representative examples of each recommendation type for these PIE contexts.

Table 3
Examples of recommendations across different Personalized Information Environments (PIEs).

Recommendation Type	Cat Video PIE	Tomato PIE
Invalid Recommendation	Talkshow clip (not an animal video)	Butter chicken recipe (not Italian)
In-PIE Recommendation	Video of a cat rolling over	Classic margherita pizza recipe
Out-PIE Recommendation	Video of a koala sleeping	Pesto pasta with pine nuts recipe

After running query tests, the number of recommendations in each buckets for different strategies will allow us to quantify the tradeoff between respecting user's stated preferences and introducing novel, out-of-distribution content in recommendations. This would allow us to assess each strategy's effectiveness in balancing personalization and recommendation diversity.

7. Reflection and Future Work

Are there any limitations in your approach?

A key limitation of our approach lies in the scalability of large language models in federated learning environments. While our experiments show that relatively small models like Qwen3-0.6B can still yield strong performance, larger models could achieve even better results in both reasoning and generalization. However, using larger models significantly narrows the pool of eligible client devices in FL, as they demand greater computational resources and memory. This presents a tradeoff between model capacity and practical deployability across a diverse, decentralized user base.

What are your planned next steps to complete your investigation?

Our immediate next steps are focused on two fronts. First, we will implement and rigorously evaluate the KG adaptation strategy to test its effectiveness in enabling LLMs to recommend outside a user's PIE without compromising relevance. Second, we will conduct a thorough differential privacy analysis, comparing the performance resilience of KTO and DPO across varying privacy budgets. These experiments will help validate **H2** and **H3** and further assess the broader potential of our KGC model to operate effectively in decentralized, privacy-preserving, and user-adaptive settings.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Gemini and Grammarly in order to rephrase some of the sentences and also to fix grammar and spelling issues. After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR, 2017.

[2] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 3580–3599. doi:10.1109/TKDE.2024.3352100.

[3] D. Nathani, J. Chauhan, C. Sharma, M. Kaul, Learning attention-based embeddings for relation prediction in knowledge graphs, arXiv preprint arXiv:1906.01195 (2019).

[4] Z. Zhang, J. Cai, Y. Zhang, J. Wang, Learning hierarchy-aware knowledge graph embeddings for link prediction, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 3065–3072.

[5] M. Chen, W. Zhang, Z. Yuan, Y. Jia, H. Chen, Fede: Embedding knowledge graphs in federated setting, in: Proceedings of the 10th International Joint Conference on Knowledge Graphs, 2021, pp. 80–88.

[6] R. Ye, W. Wang, J. Chai, D. Li, Z. Li, Y. Xu, Y. Du, Y. Wang, S. Chen, Openfedllm: Training large language models on decentralized private data via federated learning, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6137–6147.

[7] R. Ye, R. Ge, X. Zhu, J. Chai, Y. Du, Y. Liu, Y. Wang, S. Chen, FedLLM-Bench: Realistic benchmarks for federated learning of large language models, arXiv preprint arXiv:2406.04845 (2024).

[8] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, C. Finn, Direct preference optimization: your language model is secretly a reward model, 2023.

[9] L.-P. Meyer, C. Stadler, J. Frey, N. Radtke, K. Junghanns, R. Meissner, G. Dziwis, K. Bulert, M. Martin, Llm-assisted knowledge graph engineering: Experiments with chatgpt, in: Working conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow, Springer Fachmedien Wiesbaden Wiesbaden, 2023, pp. 103–115.

[10] Z. Qiu, L. Luo, S. Pan, A. W.-C. Liew, Unveiling user preferences: A knowledge graph and llm-driven approach for conversational recommendation, arXiv preprint arXiv:2411.14459 (2024).

[11] T. Donkers, J. Ziegler, The dual echo chamber: Modeling social media polarization for interventional recommending, in: Proceedings of the 15th ACM conference on recommender systems, 2021, pp. 12–22.

[12] V. Anand, M. Yang, Z. Zhao, Mitigating filter bubbles within deep recommender systems, arXiv preprint arXiv:2209.08180 (2022).

[13] K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, D. Kiela, KTO: Model alignment as prospect theoretic optimization, 2024.

[14] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3, Springer, 2006, pp. 265–284.

[15] A. Saeidi, S. Verma, M. N. Uddin, C. Baral, Insights into alignment: Evaluating dpo and its variants across multiple tasks, arXiv preprint arXiv:2404.14723 (2024).

[16] F. Spadea, O. Seneviratne, Federated fine-tuning of large language models: Kahneman-Tversky vs. direct preference optimization, arXiv preprint arXiv:2502.14187 (2025).

[17] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following LLaMA model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[18] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in Neural Information Processing Systems 36 (2023) 46595–46623.

[19] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al., Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, See https://vicuna. lmsys. org (accessed 14 April 2023) 2 (2023) 6.

[20] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, M. Fredrikson, Universal and transferable adversarial attacks on aligned language models, 2023. URL: https://arxiv.org/abs/2307.15043. arXiv:2307.15043.

[21] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, 2020.

[22] R. Li, S. Ebrahimi Kahou, H. Schulz, V. Michalski, L. Charlin, C. Pal, Towards deep conversational recommendations, Advances in neural information processing systems 31 (2018).

[23] Hugging Face, ReDial dataset, https://huggingface.co/datasets/community-datasets/re_dial, 2024.

[24] Hugging Face, Qwen3-0.6b, https://huggingface.co/Qwen/Qwen3-0.6B, 2024.