Using Knowledge Graph Embeddings to Solve Reversal Curse

Astrid Luo, Grace Yuan, Jiaxu Liu

Dec 10, 2024

1 Introduction

When a human learns the fact "George Washington was the first president of the United States," they can naturally generalize this knowledge and correctly answer the reverse query, "Who was the first president of the United States?" This form of logical reasoning is so intuitive and self-evident to humans that it often goes unnoticed. However, autoregressive language models, which predict text based on the preceding context, frequently struggle to replicate this bidirectional reasoning. This limitation is referred to as the **Reversal Curse**, which is a fairly new issue brought up in May 2024 by Lukas Berglund et al. [1].

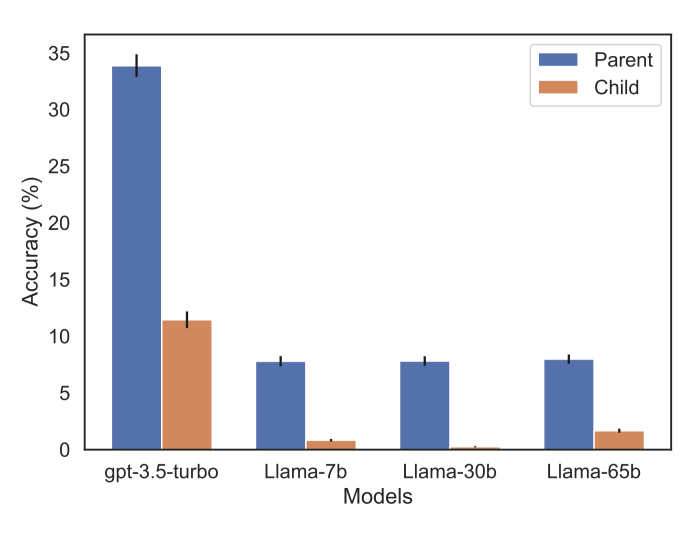

Figure 1: Accuracy of LLMs in forward (Parent) vs. reverse (Child) relationship queries, showing the Reversal Curse.

Even state-of-the-art large language models (LLMs), such as ChatGPT-4, frequently fail to answer such reverse queries. For instance, in the *celebrity-parent-child* dataset, LLMs often perform well in answering questions in the forward direction (e.g., identifying "Mary Lee Pfeiffer is Tom Cruise's mother") but perform pooly when asked the reverse (e.g., "Who is Mary Lee Pfeiffer's son?"). This asymmetry highlights a fundamental weakness in LLMs' reasoning capabilities: while they excel at memorizing and predicting statistical relationships, they struggle to infer the reverse without explicit training on both directions [1]. Instead, LLMs rely heavily on statistical patterns in their training text, rather than truly understanding or internalizing logical, bi-directional relationships.

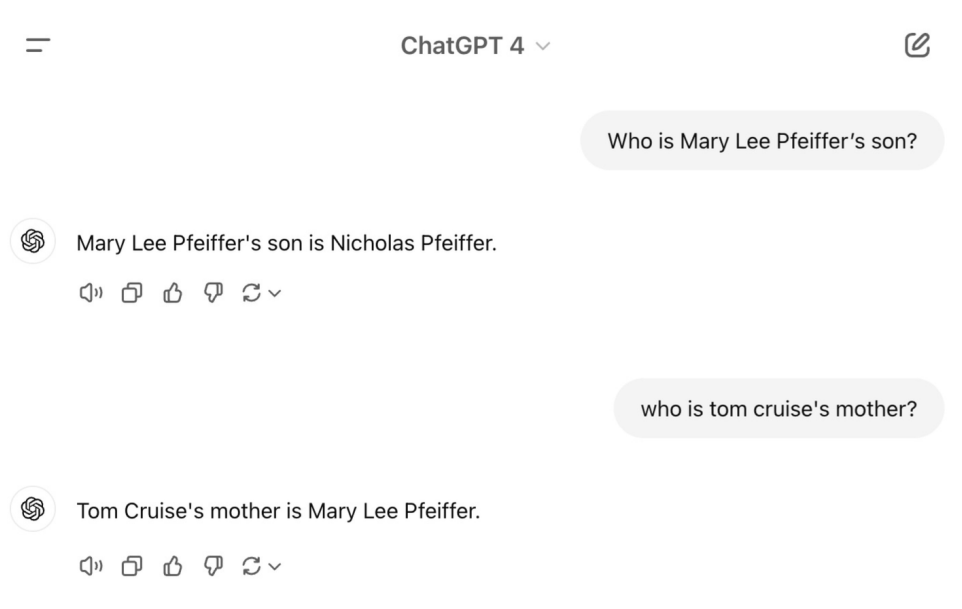

Figure 2: ChatGPT-4 fails to address Reversal Curse

The parent-child relationship is not an isolated case. The GSM8k reverse reasoning task evaluates LLMs on "backward" reasoning by testing them with math problems derived from the GSM8k dataset, requiring the model to solve problems in reverse from their training examples [6]. Similarly, in translation, models trained on one direction (e.g., Chinese-to-English) are tested in the reverse direction (e.g., English-to-Chinese), revealing limitations in reasoning due to lack of reverse-trained data [6].

Prior research has begun to document and investigate this issue. The first clear identification of the Reversal Curse showed that even highly capable models trained on "A is B" facts fail to correctly infer "B is A" [6]. Subsequent studies have linked this failure to the inherent limitations of autoregressive training, which constrains model to predict tokens based only on preceding text and leaves little room for symmetrical inference. Current mitigation attempts, such as Question-Answer (QA) fine-tuning and Chain-of-Thought (CoT) Prompting, offer limited improvements. This highlights the need for novel training approaches to improve LLM's bidirectional reasoning.

More recent efforts have focused on reshaping the training paradigm. For instance, Lin et al. [5] found that inherent model biases often caues LLMs to prioritize direct entity recall over descriptive reasoning, thereby inhibiting their ability to understand inverse relationships. Guo et al. [4] introduced the concept of Semantic-aware Permutation Training, which reorganizes source data to encourage LLMs to handle inverted relationship better. Similarly, Golovena et al. [3] introduced the concept of Semantic-aware Permutation Training which reverses the order of words in the training data while preserving certain semantic units. Another approach is **BICO** (Bidirectional Causal Language Modeling Optimization) [7], which explicitly trains LLMs on both forward and reverse contexts. However, while effective, this method often comes with significant computational overhead and does not fully resolve the underlying reasoning deficiencies. Additionally, it requires architectural change and fine-tuning of the base model, both of which demand significant computational resources.

To more comprehensively address the Reversal Curse, our approach emphasizes enhancing bidirectional reasoning through both architectural and methodological innovations, moving beyond a purely data-centric strategy and employing less computational resources. Our main contributions are as follows:

1. Developing LLMs that can reason symmetrically and effectively infer reverse relationships across various tasks via exploring alternative training methodologies and embedding strategies.
2. Advising a way to improve LLM's bidirectional reasoning that is independent to the architecture of the base model.

2 Methodology

2.1 The TransE Knowledge Graph Embedding

Our methodology integrates structured relational embeddings from a knowledge graph into the Flan-T5 model to address the challenge of learning bidirectional relationships and mitigating the reversal curse. This is achieved through the use of the TransE framework [2], which enables the representation of multi-relational data in a way that aligns naturally with bidirectional reasoning. TransE embeddings capture multi-relational data by interpreting relationships as translations between entity embeddings in a continuous vector space. Specifically, TransE models a relationship $\mathbf{r}$ between a head entity $\mathbf{h}$ and a tail entity $\mathbf{t}$ as a translation operation such that

$$\mathbf{h} + \mathbf{r} \approx \mathbf{t},$$

where $\mathbf{h}$, $\mathbf{r}$, $\mathbf{t}$ are the vector representations of the head entity, relationship, and tail entity, respectively. For example, consider the relationship "Mary Lee Pfeiffer is the parent of Tom Cruise". In this case:

- $\mathbf{h}$ = Mary Lee Pfeiffer (head entity embedding),
- $\mathbf{r}$ = parent_of (relationship embedding),
- $\mathbf{t}$ = Tom Cruise (tail entity embedding).

The TransE framework ensures that $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$, meaning the vector representation of "Mary Lee Pfeiffer" plus the translation vector for "parent_of" should approximately equal the vector representation of "Tom Cruise". Similarly, for the inverse relationship "Tom Cruise is the child of Mary Lee Pfeiffer", the embeddings would satisfy:

$$\mathbf{t} + \mathbf{r}' \approx \mathbf{h},$$

where $\mathbf{r}'$ is the embedding for the inverse relationship "child_of".

The optimization process minimizes the distance $||\mathbf{h} + \mathbf{r} - \mathbf{t}||$ for valid triples while maximizing it for corrupted triples. This approach embeds entities and relationships into a space where relationships are represented as simple translations, making it particularly effective for encoding bidirectional relationships and reasoning over them.

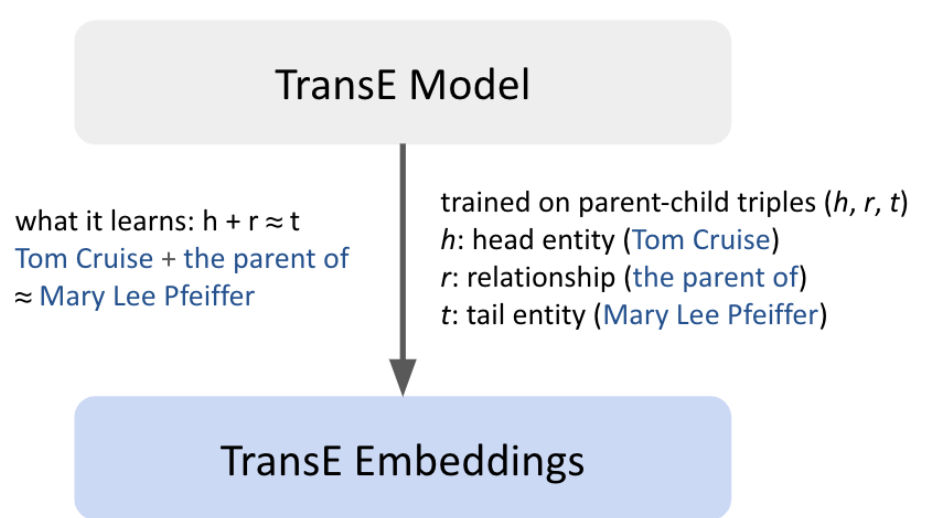

Figure 3: TransE Knowledge Graph Embedding Process

2.2 Model Architecture

To integrate TransE embeddings into our baseline Flan-T5 model, we developed a neural module that concatenates TransE-generated entity embeddings with token embeddings from Flan-T5. The TransE embeddings are first projected into the hidden dimensionality of Flan-T5 using a linear transformation. These embeddings are then appended to the sequence of token embeddings, ensuring that relational knowledge enriches the model's representation of the input.

In our implementation, we adjust the attention mask to accommodate the appended TransE embedding token, ensuring proper inclusion in the Flan-T5 encoder. This approach combines the structured relational understanding from TransE with the generative capabilities of Flan-T5, allowing the model to better capture and reason about bidirectional relationships present in the dataset. To optimize the training process, we use a learning rate scheduler that dynamically adjusts the learning rate throughout training. This ensures that the model starts with a high learning rate for rapid convergence in the initial stages and gradually reduces it to fine-tune the embeddings and the hybrid model's parameters, thereby enhancing stability and performance.

The modular design of this architecture facilitates the seamless training and inference of the hybrid model, ensuring scalable, interpretable, and efficient integration of knowledge graph embeddings.

2.3 Dataset and Prompt Design

We utilize the **celebrity relations dataset** provided by the first paper that discovered the Reversal Curse [1], which contains parent-child pairs of famous individuals. The dataset includes a total of 1,514 pairs, with structured entries specifying the names of parents, children, and the type of relationship (e.g., "mother" or "father"). This dataset provides a well-defined structure for evaluating bidirectional reasoning, as it captures explicit relationships between entities. For each parent-child pair, we generate two types of prompts:

1. **Parent-to-Child Prompt:** "Who is the child of Mary Lee Pfeiffer?"
2. **Child-to-Parent Prompt:** "Who is the parent of Tom Cruise?"

When the model is directly tested on these prompts, it performs very poorly, often failing to correctly answer most of the questions. To address this limitation and improve the baseline model's performance, we modify the prompts to include relational context. The modified prompts provide additional information to help the model understand and reason about the relationship, improving its ability to answer these queries. For example, the modified prompts are designed as:

- "Given the parent of Sasha Calle is Samira Calle, who's Samira Calle's child?"
- "Given the child of Mary Lee Pfeiffer is Tom Cruise, who's Mary Lee Pfeiffer's child?"

By enhancing the prompts with relational context, we aim to provide a more structured and interpretable input, helping the baseline model to better capture the underlying relationships in the dataset.

2.4 Further Cross-Validation and Parameter Tuning

To enhance the performance of the **FlanWithTransE** model, we explored several additional strategies during cross-validation and parameter tuning. These included applying regularization techniques, introducing a non-linear activation function (ReLU) in the neural module, and implementing a weighted embedding strategy where a linear combination of the Flan-T5-small token embeddings and the TransE embeddings was used, instead of the original concatenation method.

Regularization was applied using both L_1 and L_2 penalties, aiming to prevent overfitting. Similarly, a ReLU activation function was incorporated in the projection layer of the neural module to introduce non-linear transformations and potentially enhance the model's representational capacity. However, these modifications did not result in significant improvements in test set accuracy, with performance near the baseline achieved without these adjustments.

In contrast, the weighted embedding strategy demonstrated a notable improvement. In this approach, the final embeddings were computed as a linear combination of the Flan-T5-small token embeddings and the TransE embeddings, weighted by a tunable parameter α. Specifically, the final embedding was defined as:

$$\mathbf{E}_{\text{final}} = \alpha \cdot \mathbf{E}_{\text{FlanT5}} + (1 - \alpha) \cdot \mathbf{E}_{\text{TransE}}$$

where α controls the contribution of the two embedding types.

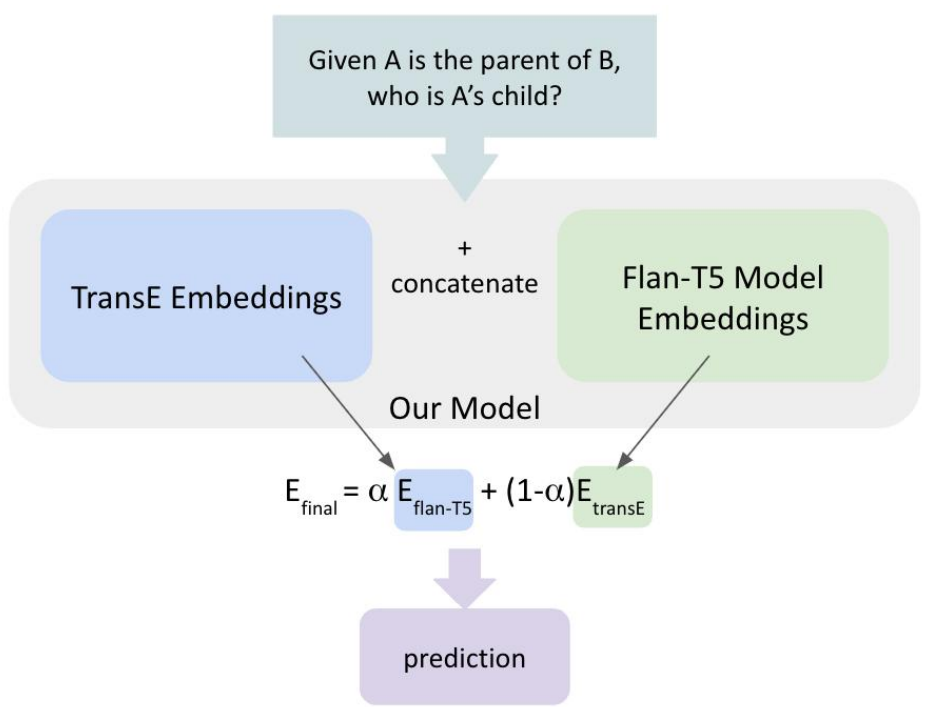

Figure 4: Workflow of integrating TransE embeddings and Flan-T5 token embeddings to form hybrid embeddings

3 Results

3.1 Baseline Model

The pretrained **Flan-T5-small** [8] served as the base model for our experiments due to its capability of handling question and answer tasks, along with constraints of limited computational resources. To evaluate its performance, we tested it on the **celebrity relations dataset**. Despite its strong generative language capabilities, the Flan-T5-small model achieved an accuracy of only **29.16%** on this dataset. This result highlights the model's limitations in capturing and reasoning about structured relational knowledge, underscoring the need for enhancements to incorporate explicit relational embeddings.

3.2 Model Training and Evaluation Across Epochs

The proposed **FlanWithTransE** model was trained on the training set and evaluated on the test set at different epochs (10, 15, 20, 25). The results demonstrated a progressive improvement in performance up to epoch 20, where the model reached its peak accuracy on the test set. Beyond epoch 20, the performance plateaued, with a slight decline observed at epoch 25, likely due to overfitting.

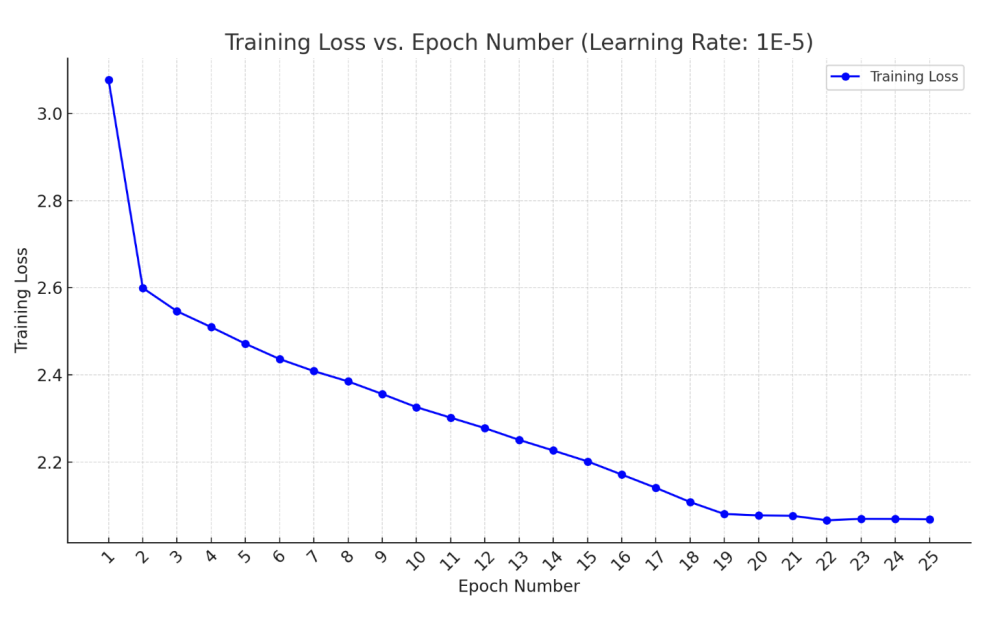

Figure 5: Training Loss at Different Epochs

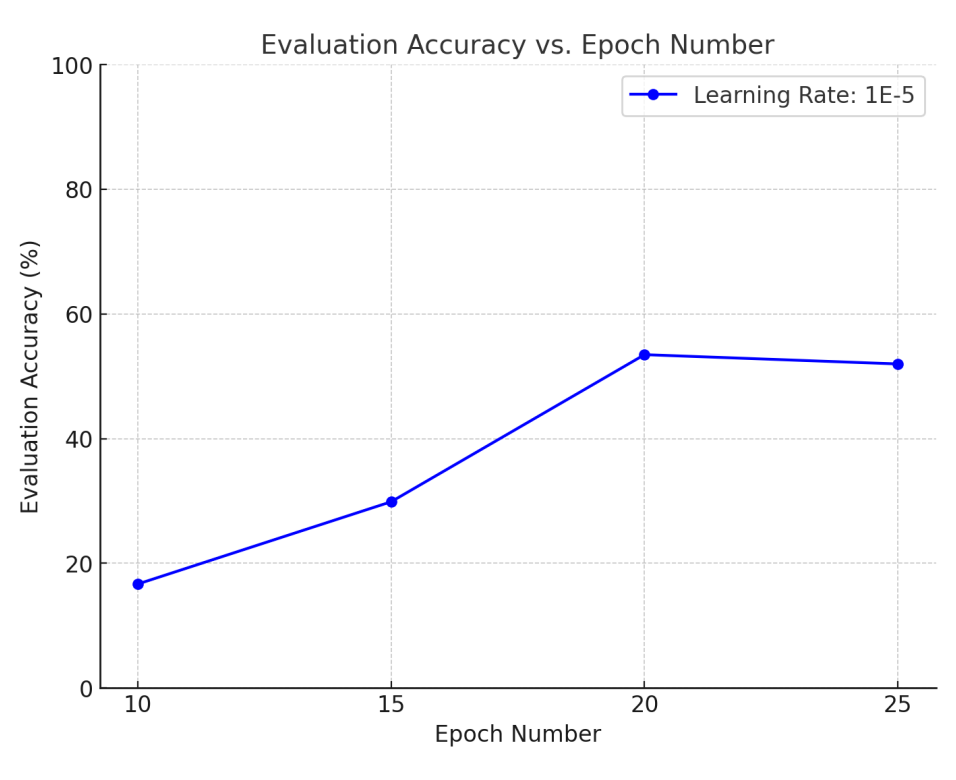

Figure 6: Evaluation Accuracy at Different Epochs

The results demonstrate that the **FlanWithTransE** architecture effectively integrates TransE embeddings with Flan-T5-small to encode relational knowledge and capture bidirectional relationships in the dataset. The model achieved its highest accuracy of **53.47%** on the test set at epoch 20 with a learning rate of 1E-5 (with a learning rate scheduler) and batch size of 32. These findings highlight the robustness of the proposed hybrid model and its capacity for enhanced relational reasoning in downstream tasks.

3.3 Parameter Tuning

Here we test the weighted embeddings method as discussed in Section 2.4. Setting alpha to 0.5 yields an evaluation accuracy of **60.4%** on the test set, approximately 10% higher than that achieved by the original method in the same number of epochs. We then tuned α across different values in $[0, 1]$ to determine the optimal balance:

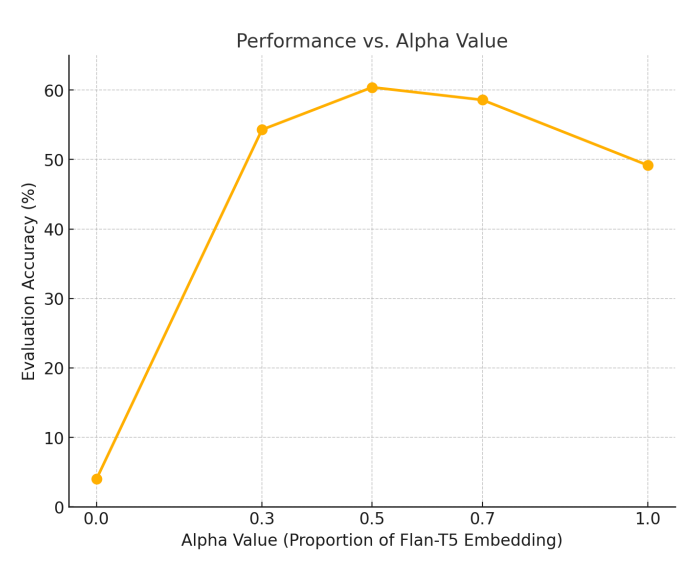

Figure 7: Evaluating the model at different alphas

The results showed that $\alpha = 0.5$ is indeed a relatively optimal setting. This indicates that a balanced mixture between the original Flan-T5-small embeddings and the translational embeddings from TransE is the most effective configuration. Additionally, setting alpha to 1, which leads to fine-tuning only FlanT5-small's original embeddings, yields test accuracy of 49.17%. It is potentially more fair to employ this result as the baseline instead, so as to address the training bias when evaluating the net gain of incorporating TransE. We further note that using pure transE embedding yields poor outcomes, this is expected as in that case we focus entirely on the translational relationships and in turn ignore grammar and semantic structures captured in the embeddings of the original model.

3.4 Summary of Findings

Compared to the baseline model's performance of **29.16%** without fine-tuning and **49.17%** after fine-tuning, our model, specifically the weighted embedding variant, achieves a peak test accuracy of **60.4%**. Therefore, the proposed **FlanWithTransE** model, which integrates Flan-T5-small token embeddings with relational TransE embeddings, significantly enhance the model's ability to learn reciprocal relationships, addressing the reversal curse in the context of personal relationships.

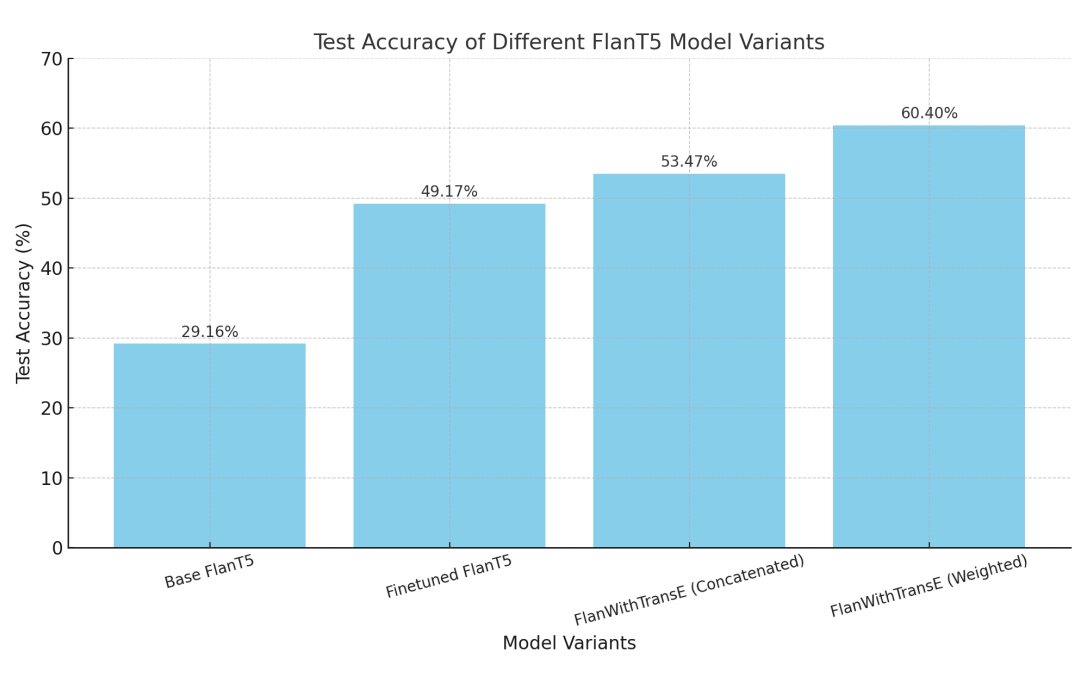

Figure 8: Comparison in performance of the baseline FlanT5 and FlanWithTransE

4 Conclusion

Our findings underscore the effectiveness of geometric relational embedding in learning bidirectional relationships and improving models' reasoning and prediction capabilities. In addition to addressing the reversal curse (in limited context), our approach features a robust framework for incorporating relational representation into existing models, paving the way for further advancements in knowledge-augmented generative modeling.

It is important to note that our approach is base model-independent: knowledge graph embeddings can be integrated into any existing LLMs without requiring fine-tuning of the models themselves.

5 Challenges and Future Work

Our model is designed to capture bidirectional reasoning across a variety of relationship contexts. As a result, while it demonstrates significant efficiency in the context of the **celebrity relations dataset**, it will not necessarily excel in tasks such as solving reverse mathematical problems. These tasks require more complex reasoning rather than simply learning straightforward relationships through knowledge graph embeddings. In such scenarios, the model's ability to generalize complex reasoning processes beyond simple relational mappings is crucial and presents an area for further development.

References

[1] Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A", 2024.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pages 2787–2795, 2013.

[3] Tatiana Golovneva, Eunji Park, Riley Smith, and Lucas Chen. Reverse training to nurse the reversal curse, 2024.

[4] Jingwen Guo, Wei Zhang, Yuhan Liu, and Meng Li. Mitigating reversal curse in large language models via semantic-aware permutation training, 2024.

[5] Tianrui Lin, Wei Huang, Yihan Zhang, Shuhao Chen, and Jia Yu. Delving into the reversal curse: How far can large language models generalize?, 2024.

[6] Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhan Chen, Ji-Rong Wen, and Rui Yan. An analysis and mitigation of the reversal curse, 2024.

[7] Zhengxin Lv, Hua Li, and Yi Zhang. An analysis and mitigation of the reversal curse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8734–8746, Singapore, December 2024. Association for Computational Linguistics.

[8] Google Research. Flan-t5-small. https://huggingface.co/google/flan-t5-small, 2022. Accessed: 2024-12-10.