ALIGNed-LLM: Adaptive Learning of Instruction-tuned Graphs and Natural Language Models

Anonymous ACL submission

Abstract

Large language models have transformed natural language processing (NLP) tasks such as question answering, dialogue generation, summarization, and so on; yet their susceptibility to hallucination remains a major challenge. Among numerous approaches to overcome this challenge, integrating Knowledge Graphs (KGs) into language models (LMs) has emerged as a promising solution, providing structured, reliable, domain-specific, and up-to-date external information. In this paper, we introduce ALIGNed-LLM, a simple yet effective approach to improve LMs' factuality by infusing KGs into their latent space using a lean strategy inspired by LLaVA. This alignment enables the LM to distinguish between similar entities, improving factual grounding and reducing hallucination. We use embeddings from a pre-trained Knowledge Graph Embedding model, such as TransE, and a trainable projection layer to align entity and text embeddings. We tested our approach on three popular KG question-answering benchmarks, alongside LMs of varying sizes, including a real-world financial use case from a large central bank in Europe that demands high accuracy and precision, demonstrating a substantial improvement in the LM's answers.

1 Introduction

The emergence of Large Language Models (LLMs) such as GPT-4 (OpenAI et al., 2024), Llama (Touvron et al., 2023), and Claude (Anthropic, 2023) has led to a step change in performance across NLP tasks, including question answering (QA), dialogue generation, and summarization. Despite these advances, factuality remains a central limitation for LLMs. Such models are constrained by stale training data (Peng et al., 2023), exhibit a persistent tendency to hallucinate (Ji et al., 2023), and often lack deep domain expertise (Zheng et al., 2024). In high-stakes domains such as finance, healthcare, and regulation, these factors can cause critical issues. While hallucinations are to some extent inherent to current LLM architectures (Banerjee et al., 2024), extensive research aims to mitigate them.

A first major line of work is retrieval-augmented generation (RAG), which improves factual accuracy by grounding LLM outputs in retrieved documents or passages (Wu and Tsioutsiouliklis, 2024; Mallen et al., 2022; Baek et al., 2023). However, RAG's effectiveness requires a comprehensive, queryable text corpus or structured data that accurately captures the organization's operational knowledge. In domains such as banking and insurance, knowledge is rarely curated as NL documents, but rather as a structured, access-controlled systems such as banking platforms, pricing engines, regulatory rule bases, where information is represented as relational tables, nested JSON, or KGs. Applying RAG in these settings typically requires building a parallel layer that mirrors the underlying systems, introducing substantial engineering overhead and challenges around data freshness and latency. In this work, we explicitly base our assumption on settings where knowledge is predominantly structured, and RAG's core assumption (knowledge as text) is fundamentally misaligned with reality.

A second major line of work is Knowledge–augmented generation (KAG) where LLMs are coupled with KGs to ground their reasoning in rich, domain-specific information (Wu and Tsioutsiouliklis, 2024; Brown et al., 2020). In practice, however, these methods typically require linearizing KGs into text, which can lead to a loss of important structural information (Guo et al., 2024; Baek et al., 2023; Wang et al., 2023). Moreover, KAG pipelines often increase computational and latency costs. In particular, we focus on settings where enterprise KGs are large, tightly governed, and not text-friendly, so KAG no longer applies.

Let us consider a real-world use case from a large European central bank where the goal is

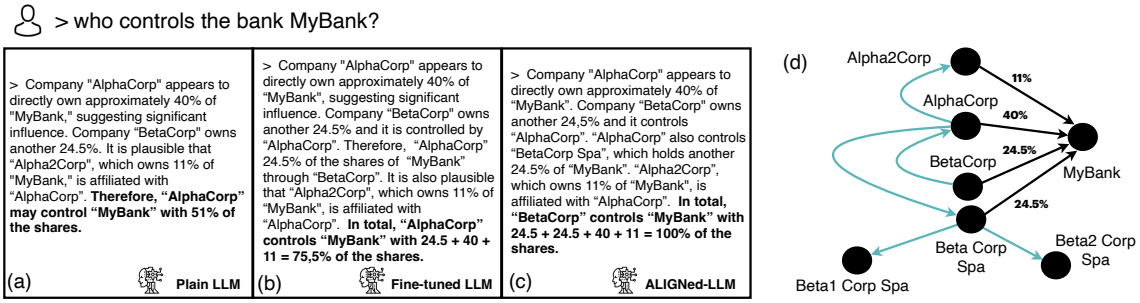

Figure 1: Three approaches to address a central bank company control QA setting. In an anonymized ownership KG fragment (d); ownership edges, weighted with the shares, are in black, while direct control edges are in blue.

to determine who ultimately controls the bank `MyBank`, according to the following recursive definition (Gulino et al., 2021): a bank, or, more in general, a company X is considered to control Company Y if X either directly owns more than 50% of Y, or indirectly does so by controlling a group of companies that collectively own more than 50% of Y. If X is not controlled by any other company, it is named as the "ultimate controller" of Y. Consider Figure 1(a). A plain LLM is used to determine the ultimate controller of `MyBank`. While the model correctly guesses that `Alpha2Corp` is a subsidiary of `AlphaCorp`, it incorrectly concludes that `AlphaCorp` controls `MyBank`, as it seems to be unaware (stale training data) that `BetaCorp` in turn controls `AlphaCorp`. In Figure 1(b), an LLM fine-tuned with KG data still incorrectly concludes that `AlphaCorp` controls `MyBank`, since the model mistakes (hallucination) `BetaCorp`, which controls `AlphaCorp`, for its subsidiary `Beta Corp Spa`. After the fine-tuning process, the model fails to disambiguate these two entities, likely due to their similar names and relationships.

In this paper, we propose a fundamentally different approach to coupling LLMs with structured knowledge for high-stakes applications, where factual accuracy is non-negotiable and knowledge primarily exists in structured form, by directly projecting structured knowledge as *Knowledge Graph Embeddings* (KGEs) into the LLM's representation space. Our approach is motivated by LLaVA (Liu et al., 2024), a simple yet effective way of bridging two different modalities where visual data and textual instruction are combined to enhance natural language QA. In particular, we leverage pretrained KGEs, i.e., encoding KG entities as low-dimensional vector spaces that capture their semantic and structural properties (Ji et al., 2022), and align them with text embeddings via a lightweight projection layer. Our infusion technique improves the ability of the LLM to resolve ambiguities and differentiate between similar entities, thanks to a rich semantic context, such as distinguishing `BetaCorp` and `Beta Corp Spa` in Figure 1(c).

Therefore, by shifting the challenge from pipeline engineering to representation learning, our method provides a natural fit for the large class of production environments where RAG's architectural assumptions are misaligned with the actual form and governance of enterprise knowledge. We summarize our main contributions as: (1) ALIGNed-LLM[1]: A projection-based architecture that directly injects the KGE into LLM's latent space removing the need for document corpora and retrieval pipelines. (2) A lightweight trainable projection layer: that aligns the embedding of the reference entities with the embedding space of the LLM. (3) Empirical validation of efficiency and accuracy: Our approach enables a 1B-parameter model to substantially outperform 13B-parameter model on enterprise KGs, such as the ownership KG of the central bank along with showcasing its ability to improve factual accuracy on public dataset.

2 Related Work

A novel approach to integrating KGs into LLMs is introduced by KG-Adapter (Tian et al., 2024), which encodes KG structure information by incorporating both node-centric and relation-centric data, as well as triples, using GNN and MLP. KG-Rank (Yang et al.) is designed for the medical domain and improves QA by automatically identifying entities in a query to retrieve related triples from the graph, while optimizing their ordering (ranking and reranking) to provide the most precise information at inference. ChatKBQA (Luo

[1] https://anonymous.4open.science/r/ALIGNed-LLM-4E77

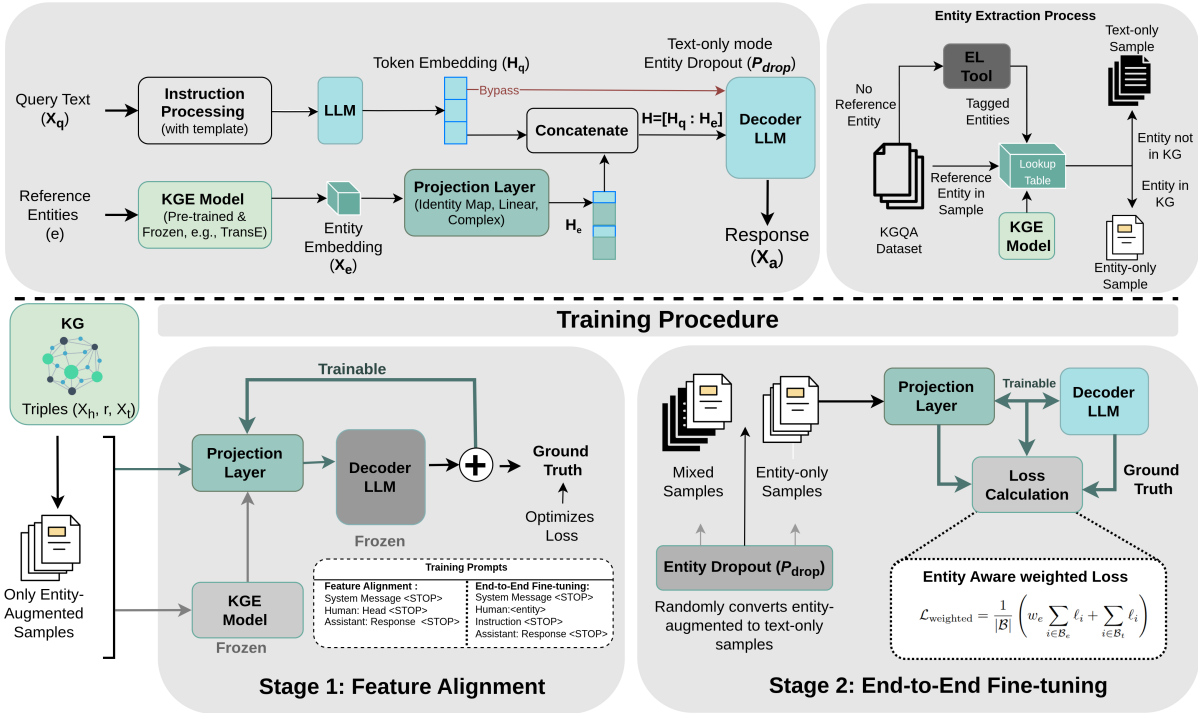

Figure 2: Architecture of ALIGNed-LLM model. Top section depicts the end to end work-flow, while the bottom section illustrates the training procedure.

et al., 2023) introduces a Generate-then-Retrieve framework, where the logical form for NL question answering is first generated using fine-tuned LLMs. Entities and relations are then retrieved and substituted through an unsupervised retrieval method. REANO (Fang et al., 2024) proposes enhancing the reader model with a retrieval-augmented knowledge graph generation module. It generates a KG from the given passages and uses a GNN to identify the top-k triples relevant to the questions.

Autonomous Knowledge Graph-based Retrofitting (KGR) (Guan et al., 2024) uses LLMs to verify information against KGs by having the LLM itself extract, detect, select, verify, and retrofit facts during the multi-step reasoning process. KnowLA (Luo et al., 2024) integrates KGE to improve Parameter-Efficient-Finetuning of LLMs by activating relevant knowledge in an input question by infusing the relevant entity embeddings extracted using an entity linking method into the respective textual token embeddings of the input text while fine-tuning the decoder by inserting a trainable adapter layer into the LLM.

While these approaches demonstrate significant achievements, they tend to be costly and poorly scalable. Moreover, they primarily focus on relevant fact retrieval, sub-graph, and small-scale graph generation techniques. Consequently, they fail to capture the complete structure and intricacies of the full KG. Additionally, these methods often require multiple fine-tuning stages and rely heavily on integrating various external tools, increasing complexity and computational overhead . Our method addresses these limitations by aligning the reference entity embedding directly with the language model embedding space while capturing the global KG structure using two-stage fine-tuning.

3 ALIGNed-LLM Approach

The primary goal of our work is to employ pre-trained KGEs to enhance the performance of LMs. Specifically, we want to improve the LLM's factual accuracy by biasing their outputs toward the facts of our domain of interest by incorporating external information from a large-scale, structured sources such as KGs. To do so, we propose ALIGNed-LLM, an Adaptive Learning of Instruction-tuned Graphs and Language Models. The pipeline of ALIGNed-LLM is shown in Figure 2. We use a pre-trained KGE model that includes entity embeddings. This model is frozen in our approach, i.e., requires no updates.

3.1 Flexible Input Processing

Entity-Augmented Mode. The QA dataset consists of pairs of reference entities and a query, pro-

cessed according to the LLM's instruction template. In our approach, we assume that the query is constructed around a set of reference entities $\mathcal{E}$. In the instruction processing step, we retrieve the entity embeddings X_e corresponding to the textual query X_q, which are then projected sequentially into the text embedding space using a lightweight trainable projection layer. After that, each transformed entity embedding H_e is concatenated with the query's language token embedding H_q. The combined input containing both textual context and entity-specific information is then passed to a decoder LM to generate the final response X_a.

Text-Only Mode: The textual query X_q , when no entity is provided, is processed through the ALIGNed-LLM's embedding layer functioning as a standard language model.

Entity Dropout: We use a special regularization technique to teach the model to consider the entity as a supplementary context rather than a mandatory signal. During Training, the entity-augmented samples are converted to text-only samples with a probability p_{drop} to ensure robust performance in both entity-augmented and text-only cases.

3.2 Generation of Entity Embedding

A Knowledge Graph is a multi-relational directed graph $\mathcal{KG} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$ where $\mathcal{E}, \mathcal{R}$ are the sets of nodes (entities) and edges (relations between entities), respectively. The set $\mathcal{T} = \{(X_h, r, X_t) \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}\}$ contains all the triples of the $\mathcal{KG}$, also called positive or true facts; h, t, and r are the head, the tail entities, and the relation, respectively.

A Knowledge Graph Embedding model learns vector representations of entities ($\mathcal{E}$) and relations ($\mathcal{R}$). A vector representation denoted by $(\mathbf{X}_h, \mathbf{r}, \mathbf{X}_t)$ is learnt by the model per triple (X_h, r, X_t), where $\mathbf{X}_h, \mathbf{X}_t \in \mathbf{V}^{d_e}$, $\mathbf{r} \in \mathbf{V}^{d_r}$. $\mathbf{V}^d$ is a d-dimensional vector space, and d_r and d_e denote relations r and entities e embedding dimensions, respectively. TransE (Bordes et al., 2013) is the first translation-based and a state-of-the-art KGE model. TransE embeds entities and relations into a single real vector space. Relationships are exclusively modeled as translations between these vectors such that $\mathbf{X_h} + \mathbf{r} \approx \mathbf{X_t}$.

TransE stands out for its analytical simplicity, making it easy to implement while ensuring computational efficiency and scalability to large datasets. Motivated by these advantages, we decided to use TransE for our study. However, TransE can be replaced by any other KGE model.

3.3 Projection Layer

The projection layer ϕ is introduced to bring entity embeddings $\mathbf{X}_e \in \mathbb{R}^{d_e}$ into the text embedding space $\mathbb{R}^{d_q}$. Thus, after retrieving the entity embedding $\mathbf{X}_e$ from the lookup table, $\mathbf{X}_e$ is passed through a trainable neural network to obtain

$$\mathbf{H}_e = \phi(\mathbf{X}_e) \in \mathbb{R}^{d_q}. \tag{1}$$

Three types of projection layers are implemented: (1) Identity Map: The projection layer is an identity map that directly embeds entities into the text embedding vector space. (2) Linear Projection: A simple linear transformation is applied on the input, providing linear mapping of the entity embedding into the text embedding space. (3) Complex Projection: It incorporates a multi-layer perceptron (MLP) with a Gaussian Error Linear Unit (GELU) activation function (Hendrycks and Gimpel, 2016). This enables more sophisticated transformations for complex alignment between the entity and the language model's text embedding spaces.

3.4 Infusion of KGE

Once the entity and text embeddings have been aligned, the final step is to infuse both embeddings and provide them as input to the LM. For entity-augmented samples, the integration is achieved by a tensor concatenation operation, replacing entity token positions with their corresponding projected embeddings:

$$\begin{aligned}\mathbf{H} = [&\mathbf{H}_q^{(1)}, \ldots, \mathbf{H}_q^{(i-1)}, \mathbf{H}_e^{(1)}, \\ &\mathbf{H}_q^{(i+1)}, \ldots, \mathbf{H}_e^{(k)}, \ldots, \mathbf{H}_q^{(n)}]\end{aligned} \tag{2}$$

where $\mathbf{H}_e^{(1)}$, ..., $\mathbf{H}_e^{(k)}$ are the projected embeddings for k entities, inserted at their respective token positions. In text-only mode, this stage is bypassed and the same decoder LLM processes both modes. With a probabilistic formulation, we maximize the likelihood of generating the correct response X_a, for the entity-augmented mode by equations 3. L is the length of the response sequence and $\mathbf{X_a^{<l}}$ represents the token generated up to l^{th} position.

$$\begin{aligned}P(\mathbf{X_a} \mid \mathbf{X_q}, \mathbf{X_e}) &= P(\mathbf{X_a} \mid \mathbf{H}) \\ &= \prod_{l=1}^{L} P(\mathbf{X_a^l} \mid \mathbf{X_a^{<l}}, \mathbf{H})\end{aligned} \tag{3}$$

3.5 Training Procedure

The training procedure follows a two-stage process consisting of (i) a *feature alignment*, and (ii) an *end-to-end fine-tuning*. Our approach accommodates different data configurations, enabling training batches with all entity-augmented samples, all text-only samples, and mixed samples. The objective is to minimize the distance between the generated response by the language model $\mathcal{L}_{\text{LM}}$ and the ground truth, using a loss function:

$$\mathbf{X_a} = \mathcal{L}_{\text{LM}}(\mathbf{H})$$
$$\mathcal{L}_{\text{fine-tune}} = CrossEntropy(\mathbf{X_a}, GroundTruth)$$

Stage 1: Feature Alignment. In this stage, the entity and text features are aligned by training the projection layer while keeping the LM's weights and biases frozen. By training only the projection matrix on entity-augmented samples, the entity representations are effectively mapped into the model's latent space.

Stage 2: End-to-End Fine-tuning. In the second stage, the projection matrix and the LM's head layer are jointly trained. This process enhances the model's ability to generate accurate responses grounded in both text and the KG. In order to integrate dual-mode training, we employ two complementary techniques. (i) Entity Dropout: The goal is to expose the models to the same question distribution both with entity context and without entity, resulting in the model learning to see entity as a supplementary signal rather than being fully dependent on it. After a small ablation with the dropout probability, during training, the entity-augmented samples are dropped randomly with $p_{\text{drop}} = 0.15$. (ii) Entity Aware Weighted Loss: We give more weight to the entity-augmented sample by employing cost-sensitive loss weighting (Elkan, 2001)(Khan et al., 2017) for final loss calculation.

$$\mathcal{L}_{\text{weighted}} = \frac{1}{|\mathcal{B}|}\left(w_e \sum_{i \in \mathcal{B}_e} \ell_i + \sum_{i \in \mathcal{B}_t} \ell_i\right) \quad (4)$$

where $\mathcal{B}_e$ and $\mathcal{B}_t$ are the entity- and text-only subsets of batch $\mathcal{B}$, ℓ_i is the per-sample cross-entropy loss, and $w_e > 1$ is a hyperparameter that decides the emphasis given to the entity-augmented sample.

The fine-tuning process supports three batch configurations. (i) Entity-only batch where each sample contains reference entity/entities. Here, both the projection layer and the LM receive a gradient from every sample, contributing to balanced optimization. (ii) text-only batch where no sample contains a reference entity. In this case, only the LM is updated, since the projection layer does not receive gradients (it is excluded from the computation graph). (iii) Mixed-batch contains both type of samples. The LM receives gradients from both, while the projection layer receives gradients only from the entity-augmented sample. Separate learning rates are used for the projection layer and LM to account for different update frequencies.

Table 1: The number of distinct entities, relations, training and test sets, and pre-trained TransE MRR score are reported here. The embedding dimension of YAGO3-10 and CO KGs is 1024, and Wikidata5M is 512.

Dataset	$\|E\|$	$\|R\|$	Train	Test	MRR
YAGO3-10	112,981	37	101,683	11,298	0.515
Mintaka	4,893	–	11,298	3,702	0.253
Wikidata	18,006	129	22,430	633	0.253
CO	14,720	7	117,074	13,008	0.570

4 Experiments

Our experiment uses two settings: (i) *Entity-only Training*, where all training samples contain an entity — isolating the contribution of KGE integration (ALIGNed's core mechanism); and (ii) *Mixed Training*, where both entity-augmented and text-only samples are used for training—demonstrating realistic deployment and utility of KGE when available. Additional results and implementation details are provided in the Appendix.

Question Answering Task. The KGQA task, structured as tuples (X_h, Xq, X_t) (Figure 2), involves answering NL questions using factually correct and structured data. Each question is associated with a specific entity (*head entity*) and its relation to the answer (*tail entity*). Formally, the task is to predict the tail entity X_t given a question X_q and a head entity X_h using the relation r, enabling reasoning over the KG to find X_t, or vice versa.

Evaluation Metrics. The metrics adopted for the analysis are ROUGE-L, ROUGE-1, F1 score, and Exact Match (EM). We also use the BLEU metric, specifically the Reverse Weighted BLEU (RWB), which prioritizes lower-order n-grams (BLEU-1 to BLEU-4) to emphasize content coverage.

Baseline Models and Datasets. We consider three open-source LLMs for our experiments: *TinyLlama-1.1B (TL-1.1B)*(Zhang et al., 2024), *vicuna-13b-v1.5 (Vicuna-13B)*(Zheng et al., 2023), *Mistral-7B-Instruct-v0.3 (Mistral-7B)*(Jiang et al., 2023). To evaluate the exact improvement intro-

duced by ALIGNed-LLM framework, LLMs fine-tuned on the textual version of the KGQA datasets are used as a baseline. In the tables, we refer to our solution using the prefix *ALIGNed-'LLM Name'*.

We consider three KGs, and four related QA datasets, as shown in Table 1. For Wikidata5M (Wang et al., 2021) KG, we considered the following two QA datasets: **(1) Wikidata QA dataset** (Banerjee et al., 2023), and **(2) Mintaka QA dataset** (Sen et al., 2022). Using YAGO3 KG (Mahdisoltani et al., 2024), we generated the **(3) YAGO3-10 QA** dataset using a question template (see Appendix) based on 37 relations .

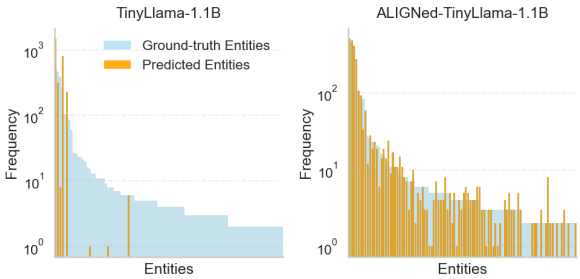

Figure 3: Distribution of most frequent entities in the answers of fine-tuned and ALIGNed-TL1.1B (CO KG).

Finally, the **(4) Company Ownership (CO)** QA dataset is constructed from an anonymized fragment of the CO KG. This KG from a central bank captures the relationships between companies: nodes represent individuals or companies, and edges are the shareholding links between them, such as ownership, control, ultimate control (as defined in Section 1), and so on. The overall graph comprises approximately 1,657M weakly connected components, and its degree distribution follows a scale-free power-law, as common in corporate economics settings. We consider a fragment of around 29,318 distinct entities, from which we extracted around 100k triples. We build the CO dataset considering two structured QA templates of increasing complexity (see Appendix), for each of the 7 types of relationships in the KG.

As for obtaining the KGE, we used the publicly accessible pre-trained TransE model (Zhu et al., 2019).To obtain the KGEs of the YAGO3-10 and CO KGs, we trained the TransE model within the RotatE framework (Sun et al., 2019).

4.1 Results and Analysis

Our experimental results show that aligning LMs with facts from KGs significantly improves the model's precision. Table 2 reports the performance

Table 2: ALIGNed-LLM performance across four QA datasets for the entity-only training setup.

Model	EM	R1	RL	RWB	F1
(A) Wikidata					
Vicuna-13B	0.657	0.698	0.698	0.708	0.661
ALIGNed-Vicuna-13B	0.664	0.704	0.706	0.725	0.669
Δ (%)	**0.70**	**0.60**	**0.80**	**1.70**	**0.80**
Mistral-7B	0.221	0.287	0.288	0.327	0.130
ALIGNed-Mistral-7B	0.605	0.647	0.646	0.679	0.596
Δ (%)	**38.40**	**36.00**	**35.80**	**35.20**	**46.60**
TL-1.1B	0.539	0.588	0.588	0.620	0.530
ALIGNed-TL-1.1B	0.652	0.693	0.693	0.779	0.658
Δ (%)	**11.30**	**10.50**	**10.50**	**15.90**	**12.80**
(B) YAGO					
Vicuna-13B	0.296	0.393	0.393	0.258	0.274
ALIGNed-Vicuna-13B	0.313	0.421	0.421	0.287	0.291
Δ (%)	**1.70**	**2.80**	**2.80**	**2.90**	**1.70**
Mistral-7B	0.000	0.031	0.031	0.008	0.000
ALIGNed-Mistral-7B	0.134	0.178	0.178	0.100	0.052
Δ (%)	**13.40**	**14.70**	**14.70**	**9.20**	**5.20**
TL-1.1B	0.253	0.353	0.352	0.228	0.227
ALIGNed-TL-1.1B	0.301	0.408	0.408	0.273	0.269
Δ (%)	**4.80**	**5.50**	**5.60**	**4.50**	**4.20**
(C) Company Ownership					
Vicuna-13B	0.648	0.666	0.666	0.381	0.623
ALIGNed-Vicuna-13B	0.713	0.730	0.730	0.484	0.690
Δ (%)	**6.50**	**6.40**	**6.40**	**10.30**	**6.70**
Mistral-7B	0.380	0.385	0.385	0.145	0.378
ALIGNed-Mistral-7B	0.488	0.511	0.512	0.235	0.473
Δ (%)	**10.80**	**12.60**	**12.70**	**9.00**	**9.50**
TL-1.1B	0.766	0.779	0.779	0.471	0.733
ALIGNed-TL-1.1B	0.901	0.916	0.915	0.789	0.897
Δ (%)	**13.50**	**13.70**	**13.60**	**31.80**	**16.40**
(D) Mintaka					
Vicuna-13B	0.576	0.627	0.627	0.586	0.567
ALIGNed-Vicuna-13B	0.630	0.677	0.678	0.638	0.623
Δ (%)	**5.40**	**5.00**	**5.10**	**5.20**	**5.60**
Mistral-7B	0.552	0.605	0.604	0.553	0.542
ALIGNed-Mistral-7B	0.568	0.618	0.617	0.587	0.557
Δ (%)	**1.60**	**1.30**	**1.30**	**3.40**	**1.50**
TL-1.1B	0.389	0.449	0.447	0.425	0.366
ALIGNed-TL-1.1B	0.402	0.462	0.462	0.430	0.380
Δ (%)	**1.30**	**1.30**	**1.50**	**0.50**	**1.40**

improvements achieved by ALIGNed-LLM over baselines across the four benchmark datasets for the entity-only training setup.

According to the EM metric, the results show a consistent and significant improvement across all datasets. For instance, on Wikidata, we observe significant Δ for ALIGNed-Mistral-7B (38.40%) and ALIGNed-TL-1.1B (11.30%), and a slight improvement for ALIGNed-Vicuna-13B (0.70%). Similarly, for the CO dataset, ALIGNed-TL-1.1B achieves a significant Δ of 13.50%, highlighting the model's capacity to improve precision in identifying exact answers compared to baseline.

ALIGNed-LLM also exhibits a consistent improvement in both the ROUGE-1 and ROUGE-L metrics. Notably, Rouge1 for the YAGO dataset improves by 14.70% (ALIGNed-Mistral-7B) and for the CO dataset by 13.70% (ALIGNed-TL-1.1B).

Table 3: Performance comparison of ALIGNed-LLM against SOTA approaches.

(A) Entity-Only Training – KGQA Datasets			
Dataset	**KnowLA**	**ALIGNed**	**Δ**
Wikidata	0.520	**0.687**	+16.7%
Mintaka	0.553	**0.674**	+12.1%
YAGO	0.330	**0.367**	+3.7%

(B) Mixed Training – Mintaka Dataset			
Model	**EM**	**Acc**	**F1**
Retrieval-Augmented Methods			
RoToR	60.0	64.5	**66.2**
KAPING	–	56.9	–
KG-Rank	–	61.9	–
Embedding-Based Methods			
KnowLA	–	50.0	–
ChatGPT-Based Methods			
KGR	53.1	–	60.8
QKR	48.0	–	54.6
CRITIC	51.0	–	58.6
CoT	53.1	–	59.3
ALIGNed (Ours)	**64.8**	**71.5**	64.5

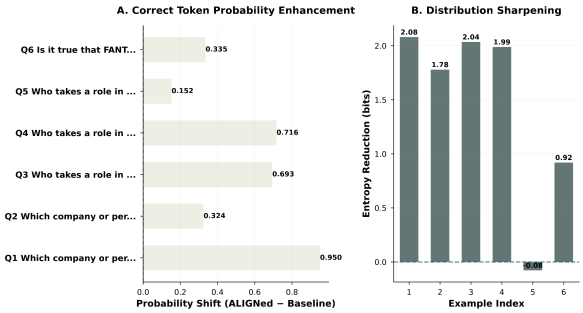

Figure 4: Analysis of probability shifts and distribution changes wrt baseline on selected 6 examples from CO test dataset. TL1.1B is used for both approaches.

For RougeL, the gains are 35.80% on the Wikidata and 14.70% on the YAGO dataset with the ALIGNed-Mistral-7B model. These results emphasize an enhanced structural alignment of the generated and reference answers.

For the RWB metric on the Wikidata, we obtain the largest relative improvement for ALIGNed-Mistral-7B (35.20%). On the CO dataset, ALIGNed-TL-1.1B and ALIGNed-Vicuna-13B achieve 31.80% and 10.30% improvements, respectively. For the F1 score, the 46.60% gain of ALIGNed-Mistral-7B on Wikidata and the 16.40% gain of ALIGNed-TL-1.1B on the CO dataset reinforce the better alignment of precision and recall of the underlying LM with our proposed approach, resulting in more accurate responses. These results provide a clear evaluation, highlighting the main advantages of ALIGNed framework over existing fine-tuning methods. Importantly, they also indicate the adaptability of ALIGNed-LLM to datasets with different entity-relation structures. Finally, we employed a T-Test considering metrics Rouge (1, 2, L, Lsum) and BLEU (1-4). Out of 24 tests, 22 were found to be statistically significant.

State-of-art Comparison. Table 3 (A) compares the ALIGNed Entity-only configuration with KnowLA, our closest baseline, across three datasets: : Wikidata, Mintaka, and YAGO. We report accuracy for our best-performing LLM and KnowLA's best public pre-trained model. To ensure fair comparison, we adopted an LLM-as-a-judge evaluation approach[2] to compute the model accuracy. Our approach clearly outperforms KnowLA across all datasets, achieving up to 16.7% performance gain on the Wikidata, demonstrating its effectiveness for integrating KGE. It is interesting to note that we also obtain a better accuracy for our model, compared to table 2(A), as the LLM-as-a-judge more accurately recognizes the correct answer (e.g., it identifies 'England' as correct when the ground truth is 'United Kingdom').

While Table 3(B) reports results comparing ALIGNed in the Mixed Training configuration on Mintaka dataset against a broader range of SOTA knowledge-augmented approaches. To ensure a fair comparison, we also report Accuracy (Acc), as some baseline methods report accuracy based on substring matching (KAPING, KnowLA). Our approach achieves the highest performance across both EM and Acc metrics, demonstrating that ALIGNed effectively leverages KGE to enhance the factual accuracy in both training configurations.

Entity-Aware Information Retrieval. In the example of Figure 1, we discussed how an LLM fine-tuned solely on KG triples may fail to disambiguate different entities with similar names (e.g., `BetaCorp` vs. `Beta Corp Spa`). By contrast, our ALIGNed models leverage projected entity embeddings to recall the correct fragment of knowledge when answering queries. Figure 3 evaluates this disambiguation capability by showing the distribution of KG entities (x-axis) across generated answers, where the y-axis reports how often each entity appears, with light blue bars denoting the ground-truth frequencies derived from reference answers.

[2]We employ *gpt-4.1-2025-04-14* model, prompted to provide a brief rationale and to check whether the predicted answer is correct with respect to the ground truth one.

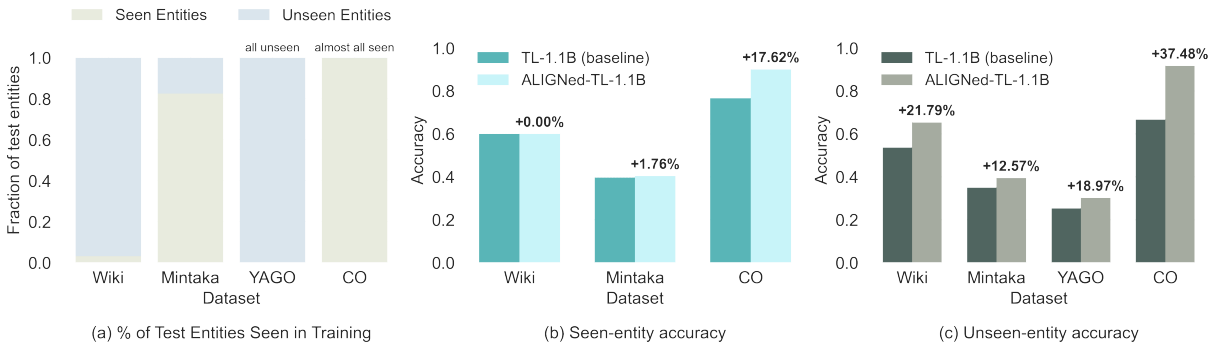

Figure 5: Generalization of ALIGNed-TL1.1B model on unseen entities from the KG.

When multiple companies share similar names, as in the CO KG, the fine-tuned baseline (left) clearly struggles to distinguish between them, with its outputs concentrating on the highest-probability entities, regardless of their correctness. Instead, our ALIGNed LLM (right), which incorporates entity embeddings alongside the input query, successfully resolves these ambiguities and produces more accurate answers with a broader coverage.

ALIGNed-LLM answer generation. Figure 4 shows how the alignment procedure shifts the generated token distribution toward facts while reducing uncertainty, across 6 heterogeneous Q/A examples from the CO KG. The difference between the correct-token probability of ALIGNed and the baseline is shown for each question in (A), where our approach consistently increases the probability. Using Shannon's entropy $\mathrm{H}(\mathrm{X}) = \sum_i p_i \log_2(p_i)$, the corresponding entropy reduction (bits) relative to the baseline is reported in (B). In 5 out of 6 examples, substantial entropy reduction suggests that ALIGNed, along with improving correct-token likelihood, produces more focused and confident predictions, highlighting the improvements stem from intelligent probability reallocation, demonstrating systematic learning of the structured knowledge.

Generalization to Unseen KG Entities. The goal of our training strategy is to maximize entity exposure to improve the model's knowledge storage, retrieval, and manipulation capabilities consistent with (Allen-Zhu and Li, 2023). Notably, the corresponding QA pairs in the test and train datasets are disjoint. Thus, genuine retrieval and reasoning, rather than memorization, are reflected by the correct response across diverse questions about the same entity, as shown in Figure 5. For questions related to both seen (exposed during training) and unseen (not seen during training but present in the KG) entities, ALIGNed maintains or improves over the baselines, proving generalization to unseen entities rather than overfitting to QA pairs.

Generalization to Multi-Entity Query. We conducted an ablation study on Mintaka, which includes comparative, intersection, and difference query types that intrinsically involve multiple entities, making it an ideal testbed for multi-entity analysis. We evaluate two configurations: (i) single-entity (primary entity embedding provided) and (ii) multi-entity (embeddings for all referenced entities injected at their respective token positions), both trained with mixed-sample procedures, achieving mean EM scores of 64.8 and 63.8, respectively. In the single-entity setup, ALIGNed outperforms baselines across all three query types (see Appendix 12), demonstrating that improving the LLM's knowledge representation supports complex multi-hop reasoning. In the multi-entity setup, performance drops by 1%, which we attribute to the difficulty of the model attending to multiple entities simultaneously and the inability of the projection layer to distinguish their respective roles.

5 Conclusion

We propose ALIGNed-LLM, a framework that infuses KGE into LLMs by aligning entity and text embeddings via a trainable projection layer, thus bridging the gap between probabilistic inference and grounded knowledge. Through extensive experiments, we validate the effectiveness of our approach in enhancing the factual accuracy and reasoning of LLMs across several KGQA datasets, reducing hallucinations, entity ambiguity, and knowledge gaps. Its potential for practical use is highlighted by its notable adaptability to domain-specific knowledge and ease of knowledge update (requiring only embedding updates). ALIGNed-LLM will be broadened to handle temporal and multi-modal KGs expanding its applicability.

Limitations

We discuss three limitations of our proposed method for establishing transparency and future research directions. ALIGNed-LLM depends on pre-trained embeddings derived from the KG, so it cannot be applied to domains where relevant embeddings or KGs are missing. We argue that similar constraints exist for all categories of RAG methods. We emphasize that this limitation is a deliberately made design choice to address resource constraints. Second, the model requires encountering sufficient entities during training to learn effective attention over the projected embeddings, as evidenced by the high performance of the small 1B model on a smaller, closed-domain KG like CO. In contrast, since the training set contains only a fraction of the entity space, performance on large-scale KGs like Wikidata is comparatively lower. This suggests that ALIGNed's performance is proportional to the ratio of training entities to the total KG entities. The third limitation is the single-entity assumption of our design, in which a single entity grounds the LLM's response. Even though our framework supports integration of multiple entities, the algorithm is not optimized for this—addressing this is a priority for our future work.

Ethical Statement

The CO dataset was processed in compliance with ethical standards and relevant laws, including data protection regulations (e.g., GDPR). All data were anonymized and analyzed within secure systems. No personally identifiable information was accessed or disclosed.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.

Anthropic. 2023. Claude: An AI Assistant. Accessed: 2025-1-1.

Jinheon Baek, Soyeong Jeong, Minki Kang, and 1 others. 2023. Knowledge-augmented language model verification. *arXiv preprint arXiv:2310.12836*.

Debayan Banerjee, Pranav Ajit Nair, Ricardo Usbeck, and Chris Biemann. 2023. Gett-qa: Graph embedding based t2t transformer for knowledge graph question answering. In *European Semantic Web Conference*, pages 279–297. Springer.

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Tom B. Brown, Benjamin Mann, Nick Ryder, and 1 others. 2020. Language models are few-shot learners. In *NeurIPS*.

Charles Elkan. 2001. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd.

Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald. 2024. Reano: Optimising retrieval-augmented reader models through knowledge graph generation. In *ACL*, pages 2094–2112.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18126–18134.

Andrea Gulino, Stefano Ceri, Georg Gottlob, Emanuel Sallinger, and Luigi Bellomarini. 2021. Distributed company control in company shareholding graphs. In *ICDE*, pages 2637–2648.

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.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv*.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Networks Learn. Syst.*, 33(2):494–514.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. 2017. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Haoran Luo, Zichen Tang, Shiyao Peng, Yikai Guo, Wentai Zhang, Ma, and 1 others. 2023. Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. *arXiv preprint arXiv:2310.08975*.

Xindi Luo, Zequn Sun, Jing Zhao, Zhe Zhao, and Wei Hu. 2024. Knowla: Enhancing parameter-efficient finetuning with knowledgeable adaptation. *arXiv preprint arXiv:2403.14950*.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian Martin Suchanek. 2024. Dataset: Yago3-10 (yet another great ontology 3-10).

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 7.

Josh Achiam OpenAI, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Almeida, and 1 others. 2024. Gpt-4 technical report, 2024. *URL https://arxiv. org/abs/2303.08774*, 2:6.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and 1 others. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. *arXiv preprint arXiv:2210.01613*.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Shiyu Tian, Yangyang Luo, Tianze Xu, Caixia Yuan, Huixing Jiang, Chen Wei, and Xiaojie Wang. 2024. Kg-adapter: Enabling knowledge graph integration in large language models through parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3813–3828.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Liu, and 1 others. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *TACL*, 9:176–194.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

Xue Wu and Kostas Tsioutsiouliklis. 2024. Thinking with knowledge graphs: Enhancing llm reasoning through structured data. *arXiv preprint arXiv:2412.10654*.

R Yang, H Liu, E Marrese-Taylor, Q Zeng, YH Ke, W Li, L Cheng, Q Chen, J Caverlee, Y Matsuo, and 1 others. Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. arxiv 2024. *arXiv preprint arXiv:2403.05881*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Jiawei Zheng, Hanghai Hong, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *arXiv preprint arXiv:2402.15061*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Zhaocheng Zhu, Shizhen Xu, Meng Qu, and Jian Tang. 2019. Graphvite: A high-performance cpu-gpu hybrid system for node embedding. In *The World Wide Web Conference*, pages 2494–2504. ACM.

A Experiment Details

A.1 Question-Answering dataset

The anonymized company ownership question-answering dataset is built according the following two templates in Table 4 and Table 5, which consists of increasingly complexity queries. For each query we build one positive example, i.e., consists of <query, correct answer>, and a negative example, i.e., <query, wrong answer>. We removed entities where the answers containing more than 20 entities from the dataset, to avoid bottlenecks and limitations for the used LLMs.

The YAGO3-10 QA dataset is constructed using the query template shown in Table 6. The queries we formulated considering the 37 relationships in

Category	Query Templates
Ultimate Control	• Who is the ultimate controller of {Y}? • Who holds ultimate control over {Y}? • Which company or person is the ultimate controller of {Y}? • Is it true that {X} is the ultimate controller of {Y}? • Does {X} have ultimate control over {Y}? • Can we confirm that {X} is the top controller of {Y}?
Control	• Who controls {Y}? • Who is the controller of {Y}? • Which company or person has control over {Y}? • Does {X} control {Y}? • Is it true that {X} has control over {Y}? • Can we verify that {X} is a controller of {Y}?
Ownership	• Who owns {Y}? • Who is the owner of {Y}? • Which entity is the owner of {Y}? • Is it true that {X} owns shares of {Y}? • Does {X} have ownership over {Y}? • Can we confirm that {X} owns shares in {Y}?
Role	• Who has a role in {Y}? • Who takes a role in {Y}? • Which company or person assumes a role with {Y}? • Does {X} have a role in {Y}? • Is it true that {X} is assigned a role in {Y}? • Can we confirm that {X} plays a role in {Y}?
Qualified Holdings	• Who has qualified holdings in {Y}? • Who is a qualified holder of {Y}? • Which company or person holds a qualified position in {Y}? • Is it true that {X} has qualified holdings in {Y}? • Does {X} qualify as a holder in {Y}? • Can we confirm that {X} is a qualified holder of {Y}?
Reachability	• Is it true that {X} is connected to {Y}? • Can {X} reach {Y} through any connections? • Does {X} have a connection to {Y}? • Is it possible for {X} to be accessed by {Y}?
Influence	• Who influences {Y}? • Who has influence over {Y}? • Is it true that {X} influences {Y}? • Does {X} have influence on {Y}? • Can {X} influence {Y}? • Is there evidence that {X} exerts influence over {Y}?

Table 4: Company Ownership Q/A.

Category	Query Templates
Control	• How many companies and/or people control {Y}? • Is it true that {Y} is controlled by {Z} companies and/or people?
Ownership	• How many companies own {Y}? • Is it true that {Y} is owned by {Z} companies and/or people?
Role	• How many entities have a role in {Y}? • Is it true that {Y} has {Z} entities assuming a role in it?
Qualified Holdings	• How many entities have qualified holdings in {Y}? • Is it true that {Y} has {Z} entities with qualified holdings?
Reachability	• How many entities can reach {Y} through any connections? • Is it true that {Y} is connected with {Z} companies and/or people?
Influence	• How many companies and/or people influence {Y}? • Is it true that {Y} is influenced by {Z} companies and/or people?

Table 5: Company Ownership reasoning Q/A.

Table 6: YAGO3-10 QA Query template.

Relation	Query Templates
wasBornIn	Where was {x} born?
worksAt	Where does {x} work? {x} is employed at which organization?
happenedIn	In which location did {x} occur?
isLeaderOf	Who/what is led by {x}?
hasWonPrize	Which prize has {x} won?
isAffiliatedTo	With whom is {x} affiliated?
hasChild	Who is a child of {x}? Which person is {x} a parent to?
imports	Which items are imported by {x}?
exports	Which items are exported by {x}?
isLocatedIn	Where is {x} located?
hasCurrency	What is the currency of {x}?
created	What was created by {x}?
graduatedFrom	From which institution did {x} graduate? {x} is an alumnus of which university/school?
actedIn	{x} acted in which movie/show? Which film/show featured {x}?
hasGender	What is the gender of {x}? {x} identifies as which gender?
hasCapital	What is the capital of {x}?
isInterestedIn	What is {x} interested in?
isConnectedTo	What is {x} connected to?
influences	Who/what is influenced by {x}?
owns	What is owned by {x}?
isKnownFor	What is {x} known for? {x} is famous for what?
directed	What did {x} direct?
diedIn	Where did {x} die?
dealsWith	What does {x} deal with?
livesIn	Where does {x} live?
isPoliticianOf	Which country/region is {x} a politician of? {x} is politically active in which place?
hasWebsite	{x} has which website?
hasNeighbor	What is the neighboring country of {x}? Which country shares a border with {x}?
hasOfficialLanguage	What is the official language of {x}?
isMarriedTo	Who is {x} married to? {x} is the spouse of whom?
hasAcademicAdvisor	Who is the academic advisor of {x}?
isCitizenOf	Which country is {x} a citizen of?
edited	Which work did {x} edit?
playsFor	Which team does {x} play for? {x} is a member of which sports team?
participatedIn	In which war/battle/event did {x} participate?
wroteMusicFor	Which production did {x} write music for? {x} composed music for which work?
hasMusicalRole	What is the musical role of {x}?

YAGO3-10 KG. For each relation group, the corresponding {x} value is replaced by the respective entities in the triples.

A.2 Implementation Details.

For all models and datasets, we employed a learning rate of 2e-4 for overall model training and 2e-5 for projection layer training. Additionally to optimize the learning process, we adopted a warm_up ratio of 0.03 to gradually adjust the learning rate. For Wikidata dataset we used 100 epoch to train the projection layer whereas for YAGO-310, Mintaka, and Company Ownership dataset we used 50 epoch to train the projection layer. On the other hand, for end-to-end fine-tuning, we used a single epoch across all models and datasets to avoid the risk

of over-fitting and knowledge degradation. Moreover, for fine-tuning the combination of LoRA and PEFT technique was used. We set $p_{\text{drop}} = 0.15$ and $w_e = 2$ to obtain the best performance on the Mixed-sample setup on the Mintaka dataset. All experiments were conducted on NVIDIA A100-SXM4 Tensor Core-GPUs with 40 GB HBM2 memory and H100-SXM5 Tensor Core-GPUs with 94 GB HBM2e memory. During training and inference, the maximum GPU usage was 64GB GPU, and the system utilized up to 128 GB of RAM.

B Result and Analysis

Tables 7-10 report detailed results of the experiments conducted on ALIGNed-LLM using 3 LLMs on 4 datasets. They also consist of a comparison study with the fine-tuned baselines on the same datasets for a fair comparison. We evaluated our model using multiple metrics to verify the consistency of our results. Table 7 demonstrates consistent performance improvement of our model on the Wikidata QA dataset for all three LLMs. Similar performance is also observed on the YAGO3-10 QA dataset in Table 8. Except for the BLEU-4 metric, improved performance is also observed on the Company Ownership QA dataset in Table 9. On the contrary, we observe mixed results on the Mintaka dataset in Table 10 indicating smaller models required further optimization on complex datasets.

B.1 Effect of Entity Embedding dimension

. Additionally, we investigate how the dimension of the entity embeddings impacts the transfer of factual knowledge into a language model. In other words, we want to know whether the preservation of more features of KG can improve the overall performance of our model. The results in Table 11 show that, except for the weighted BLEU score, all other metrics gain positively for dimension size 1024 showing better performance for higher dimensions. Figure 6 depicts the sampled training loss trend for the alignment feature training stage for training with these two different dimensions indicating higher loss values for 1024 dimension throughout the training steps.

B.2 Correlation of Data and Measurements

Although BLEU measures the challenge of aligning increasingly long phrasing, its effectiveness varies on the basis of the complexity of the alignment being measured. BLEU-1, which emphasizes

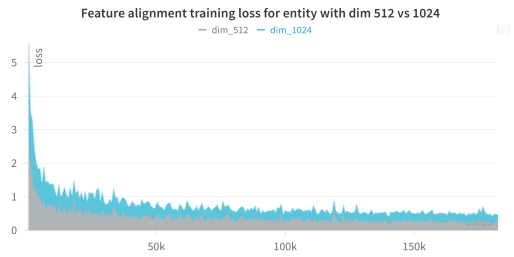

Figure 6: Feature alignment training loss of ALIGNed TL-1.1B model over CO dataset for dimension 512 and 1024

unigram precision, often scores higher as it captures individual word matches; BLEU-4, requiring longer n-gram alignment, is more sensitive to structural coherence, sentence complexity, and sequence length. Similar considerations can be developed for the ROUGE scores. In Figure 7, we analyzed this relationship in terms of the complexity and average word count of the answers across datasets.

The high textual complexity of CO causes a visible drop in BLEU-2, and only a moderate drop in BLEU-3 and BLEU-4, with good average BLEU and ROUGE scores. Less complex cases, like Wikidata are less affected by this phenomenon. Finally, despite the high complexity, with Mintaka, high BLEU scores are observed, with a drop in BLEU-4 that is mitigated by the low word count and complexity of the dataset, as discussed earlier.

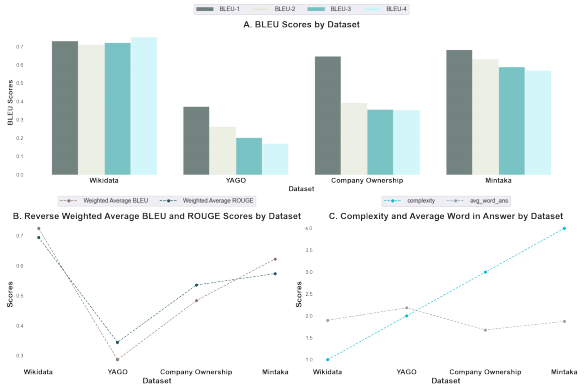

Figure 7: Evaluation metric analysis of ALIGNed-Vicuna-13B model on all datasets.

B.3 Effect of Projection Type

. In Figure 8, we show the difference of using a linear vs. a complex projection on ALIGNed-TL-1.1B model. We chose EM metric to showcase the impact. The model gains better results on all but CO dataset, when using the complex projection

Table 7: ALIGNed-LLM performance on Wikidata QA dataset

Model	F1	Rouge1	Rouge2	RougeL	RougeLsum	BLEU-1	BLEU-2	BLEU-3	BLEU-4	EM
Vicuna-13B	0.567	0.661	0.327	0.698	0.698	0.716	0.689	0.707	0.734	0.657
ALIGNed_Vicuna-13B	0.669	0.704	0.337	0.706	0.706	0.731	0.709	0.723	0.752	0.664
Gain	**0.80%**	**0.60%**	**1.0%**	**0.80%**	**0.80%**	**1.50%**	**2.00%**	**1.60%**	**1.80%**	**0.70%**
Mistral-7B	0.130	0.287	0.171	0.288	0.289	0.333	0.319	0.330	0.320	0.221
ALIGNed_Mistral-7B	0.596	0.647	0.324	0.646	0.647	0.675	0.660	0.692	0.721	0.605
Gain	**46.6%**	**36.0%**	**15.3%**	**35.8%**	**35.8%**	**34.2%**	**34.1%**	**36.2%**	**40.1%**	**38.4%**
TL-1.1B	0.530	0.588	0.292	0.588	0.588	0.620	0.603	0.635	0.644	0.539
ALIGNed_TL-1.1B	0.658	0.693	0.338	0.693	0.692	0.743	0.756	0.834	0.882	0.652
Gain	**12.8%**	**10.5%**	**4.6%**	**10.5%**	**10.4%**	**12.3%**	**15.3%**	**19.9%**	**23.8%**	**11.3%**

Table 8: ALIGNed-LLM performance on YAGO310 QA Dataset

Model	F1	Rouge1	Rouge2	RougeL	RougeLsum	BLEU-1	BLEU-2	BLEU-3	BLEU-4	EM
Vicuna-13B	0.274	0.393	0.149	0.393	0.392	0.343	0.232	0.182	0.152	0.296
ALIGNed_Vicuna-13B	0.291	0.421	0.169	0.421	0.421	0.374	0.263	0.204	0.172	0.313
Gain	**1.70%**	**2.80%**	**2.0%**	**2.80%**	**2.90%**	**3.10%**	**3.10%**	**2.2%**	**2.00%**	**1.70%**
Mistral-7B	0.000	0.031	0.002	0.031	0.031	0.019	0.001	0.000	0.000	0.000
ALIGNed_Mistral-7B	0.052	0.178	0.065	0.178	0.177	0.153	0.082	0.051	0.048	0.134
Gain	**5.2%**	**14.7%**	**6.3%**	**14.7%**	**14.6%**	**13.4%**	**8.1%**	**5.1%**	**4.8%**	**13.4%**
TL-1.1B	0.227	0.353	0.119	0.352	0.353	0.308	0.200	0.156	0.134	0.253
ALIGNed_TL-1.1B	0.269	0.408	0.153	0.408	0.407	0.363	0.249	0.188	0.158	0.301
Gain	**4.2%**	**5.5%**	**3.4%**	**5.6%**	**5.4%**	**5.5%**	**4.9%**	**3.2%**	**2.4%**	**4.8%**

Table 9: ALIGNed-LLM performance on Company Ownership QA Dataset. Notably complex projection type is used for all except ALIGNed_TL-1.1B on CO dataset where a linear projection type is used to achieve the best results.

Model	F1	Rouge1	Rouge2	RougeL	RougeLsum	BLEU-1	BLEU-2	BLEU-3	BLEU-4	EM
Vicuna-13B	0.623	0.666	0.042	0.666	0.666	0.597	0.233	0.179	0.367	0.648
ALIGNed_Vicuna-13B	0.690	0.730	0.084	0.730	0.730	0.647	0.395	0.358	0.353	0.713
Gain	**6.7%**	**6.4%**	**4.2%**	**6.4%**	**6.4%**	**5.0%**	**16.2%**	**17.9%**	**-1.4%**	**6.5%**
Mistral-7B	0.378	0.385	0.000	0.385	0.385	0.310	0.045	0.031	0.012	0.380
ALIGNed_Mistral-7B	0.473	0.511	0.020	0.512	0.511	0.436	0.125	0.077	0.076	0.488
Gain	**9.5%**	**12.6%**	**2.0%**	**12.7%**	**12.6%**	**12.6%**	**8.0%**	**4.6%**	**6.4%**	**10.8%**
TL-1.1B	0.733	0.779	0.062	0.779	0.778	0.675	0.332	0.337	0.337	0.766
ALIGNed_TL-1.1B	0.897	0.916	0.174	0.915	0.915	0.860	0.770	0.726	0.682	0.901
Gain	**16.4%**	**13.7%**	**11.2%**	**13.6%**	**13.7%**	**18.5%**	**43.8%**	**38.9%**	**34.5%**	**13.5%**

Table 10: ALIGNed-LLM performance on Mintaka dataset

Model	F1	Rouge1	Rouge2	RougeL	RougeLsum	BLEU-1	BLEU-2	BLEU-3	BLEU-4	EM
Vicuna-13B	0.567	0.627	0.301	0.627	0.627	0.638	0.583	0.531	0.499	0.576
ALIGNed_Vicuna-13B	0.623	0.677	0.331	0.678	0.678	0.684	0.632	0.590	0.570	0.630
Gain	**5.60%**	**5.00%**	**3.00%**	**5.10%**	**4.60%**	**4.90%**	**5.90%**	**7.10%**	**5.40%**	**5.20%**
Mistral-7B	0.542	0.605	0.287	0.604	0.604	0.609	0.550	0.498	0.454	0.552
ALIGNed_Mistral-7B	0.557	0.618	0.294	0.617	0.617	0.630	0.578	0.544	0.527	0.568
Gain	**1.50%**	**1.30%**	**0.70%**	**1.30%**	**1.30%**	**2.10%**	**2.80%**	**4.60%**	**7.30%**	**1.60%**
TL-1.1B	0.366	0.449	0.210	0.447	0.448	0.475	0.417	0.374	0.347	0.389
ALIGNed_TL-1.1B	0.380	0.462	0.218	0.462	0.462	0.483	0.424	0.375	0.337	0.402
Gain	**1.40%**	**1.30%**	**0.80%**	**1.50%**	**1.40%**	**0.80%**	**0.70%**	**0.10%**	**-1.00%**	**1.30%**

Table 11: ALIGNed-TL-1.1B performance on the Company Ownership dataset with two different pre-trained embedding sizes.

Dim.	EM	Rouge1	RougeL	RWB	F1
512	0.842	0.855	0.856	0.736	0.830
1024	0.848	0.863	0.862	0.727	0.837
Δ (%)	**+0.60**	**+0.80**	**+0.60**	**-0.90**	**+0.70**

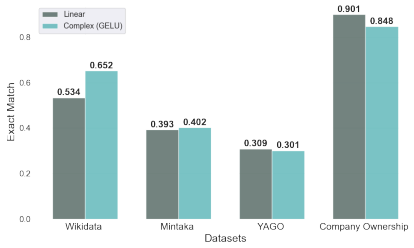

Figure 8: Linear vs. Complex Projections performance.

layer with the GELU activation function. The 5.3% improvement in the CO dataset with the linear projection vs. the complex one can be attributed to the structure of the CO KG, consisting of a less intricate relational structure. This allows the linear approach to capture the underlying pattern better without the need for a complex model. Hence, the choice of the best projection strategy may depend on the KG topology and structure.

B.4 Generalization to Multi-Entity Query

To investigate how ALIGNed handles queries that inherently involve multiple entities, we conduct an ablation study. We used the Mintaka dataset for this analysis, as it includes challenging questions such as "Which movie had a bigger budget, Avatar or Training Day?" (requiring comparison between multiple entities) and "What man was a famous American author and also a steamboat pilot on the Mississippi River?" (requiring multi-relational reasoning). The queries of type 'comparative', 'intersection', and 'difference' involve purely multiple entities, making this an ideal testbed. We evaluate ALIGNed-LLM in two configurations: (i) single-entity (primary entity embedding is provided), and (ii) multi-entity (embeddings for all referenced entities are injected at their respective token positions). For both cases, we used a mixed-sample training procedure, achieving a mean EM score of 64.8 and 63.8, respectively.

Model	Baseline	ALIGNed	Δ
Vicuna-13B	0.591	**0.640**	+4.90%
Mistral-7B	0.562	**0.580**	+1.80%
TL-1.1B	0.413	**0.436**	+2.30%

Table 12: Performance (EM Score) for Multi-entity questions on the Mintaka Dataset.

In the single-entity training configuration, ALIGNed outperforms baselines on these three query types as well, as shown in Table 12, demonstrating that improving the LLM's knowledge storage and retrieval capacity can enable more complex reasoning across multi-hop queries and multiple entities or relations. Whereas in the multi-entity training setup, the mean model performance degrades by 1% compared to the single-entity configuration. We attribute this performance drop to two facts: (i) the model struggles to appropriately attend to multiple entities simultaneously, and (ii) the incapacity of the projection layer to distinguish the roles of multiple entities. This finding reveals an important limitation of our approach and opens a promising direction for future work.

B.5 Statistical Test Results

In Table 13 and Table 14 we report a statistic significant test for BLEU and ROUGE metric, respectively. This test was conducted to verify if the performance difference between ALIGNed-LLM and fine-tuned baselines is statistically significant. We use paired t-test between our proposed model and baselines. For each metric, a significance level of 0.05 was used.

B.6 Error Analysis

The following Table 15 and Table 16 report an analysis of the QA errors, for the baseline fine-tuned TL-1.1B and the ALIGNed-TL-1.1B respectively, based on 200 examples for each model. In the tables we can observe that the baseline LLM has often more completely wrong answers, which are completely un-related to the question or the ground-truth answer. The most common type of error is answering with the most frequent entities of the anonymized QA datasets. Specifically, we identified three entities that dominate the incorrect answers:

- “ORSINI SPA", which appears in 15% and 37.5% of the answers generated by the ALIGNed-LLM and the baseline LLM, respectively.
- “SONNINO, GIANINAZZI E ARGENTI GROUP", occurring in 3% and 14.5% of the answers for the ALIGNed-LLM and the baseline LLM, respectively.
- “CIPOLLA-SANTORO E FIGLI", which is included in 3% of the answers for both models.

These entities are particularly prominent in the dataset, with approximately 10,000 occurrences each across the QA corpus. Moreover, “ORSINI SPA” frequently co-occurs with questions related to the relation "role." This error is 2 to 5 times more present in the baseline fine-tuned model, underscoring its tendency to rely on spurious associations. In contrast, the ALIGNed-LLM demonstrates superior disambiguation capabilities and a deeper understanding of the underlying data structure, reducing the prevalence of such errors.

B.7 Aligned-LLM answer generation.

For analyzing the answer generation process of ALIGNed-LLM we used six randomly selected questions shown in table 17.

Table 13: Results from Paired t-test for BLEU 1, 2, 3, and 4 scores for ALIGNed-LLM against Baselines

Dataset	Model	T-statistic	P-value	Result
CO	ALIGNed Vicuna-13B	-2.047	0.133	The results are not statistically significant (fail to reject H0)
CO	ALIGNed Mistral-7B	-4.610	0.019	The results are statistically significant (reject H0)
CO	ALIGNed TL-1.1B	-5.430	0.012	The results are statistically significant (reject H0)
YAGO3-10	ALIGNed Vicuna-13B	-8.918	0.003	The results are statistically significant (reject H0)
YAGO3-10	ALIGNed Mistral-7B	-3.936	0.029	The results are statistically significant (reject H0)
YAGO3-10	ALIGNed TL-1.1B	-5.538	0.012	The results are statistically significant (reject H0)
Wikidata	ALIGNed Vicuna-13B	-15.559	0.001	The results are statistically significant (reject H0)
Wikidata	ALIGNed Mistral-7B	-25.772	0.000	The results are statistically significant (reject H0)
Wikidata	ALIGNed TL-1.1B	-7.041	0.006	The results are statistically significant (reject H0)
Mintaka	ALIGNed Vicuna-13B	-9.960	0.002	The results are statistically significant (reject H0)
Mintaka	ALIGNed Mistral-7B	-3.621	0.036	The results are statistically significant (reject H0)
Mintaka	ALIGNed TL-1.1B	-0.363	0.741	The results are not statistically significant (fail to reject H0)

Table 14: Results from Paired t-test for ROUGE scores for ALIGNed-LLM against Baselines

Dataset	Model	T-statistic	P-value	Result
CO	ALIGNed Vicuna-13B	-10.636	0.002	The results are statistically significant (reject H0)
CO	ALIGNed Mistral-7B	-3.752	0.033	The results are statistically significant (reject H0)
CO	ALIGNed TL-1.1B	-42.333	0.000	The results are statistically significant (reject H0)
YAGO3-10	ALIGNed Vicuna-13B	-12.520	0.001	The results are statistically significant (reject H0)
YAGO3-10	ALIGNed Mistral-7B	-6.012	0.009	The results are statistically significant (reject H0)
YAGO3-10	ALIGNed TL-1.1B	-9.448	0.003	The results are statistically significant (reject H0)
Wikidata	ALIGNed Vicuna-13B	-9.798	0.002	The results are statistically significant (reject H0)
Wikidata	ALIGNed Mistral-7B	-5.975	0.009	The results are statistically significant (reject H0)
Wikidata	ALIGNed TL-1.1B	-6.136	0.009	The results are statistically significant (reject H0)
Mintaka	ALIGNed Vicuna-13B	-8.797	0.003	The results are statistically significant (reject H0)
Mintaka	ALIGNed Mistral-7B	-7.667	0.005	The results are statistically significant (reject H0)
Mintaka	ALIGNed TL-1.1B	-8.041	0.004	The results are statistically significant (reject H0)

Error Type	Percentage	Question	Ground-Truth	Predicted Answer
Completely Wrong	0.68	Who takes a role in BORSELLINO E FIGLI?	CIPOLLA-SANTORO E FIGLI	ORSINI SPA
True-False Wrong	0.28	Is it true that PETRUCCI SPA is assigned a role in GAGGINI SPA?	True	False
Subset of the Answer	0.03	Who takes a role in MANACORDA, FILANGIERI E MOSCHINO SPA?	EMO, GIACCONI E BENUSSI GROUP, ORSINI SPA	ORSINI SPA
Similar Company Type	0.02	Which entity is the owner of MARRONE-AMALDI E FIGLI?	TARCHETTI-VENDITTI GROUP	SONNINO, GIANINAZZI E ARGENTI GROUP

Table 15: Error Analysis for TL-1.1B-Chat on the anonymized CO KG

Error Type	Percentage	Question	Ground-Truth	Predicted Answer
Completely Wrong	0.49	Who takes a role in CAMPANELLA-PELLI E FIGLI?	BRICCIALDI, VIOLA E BORRANI S.R.L.	ORSINI SPA
True-False Wrong	0.43	Does VERDONE E FIGLI control PAGLIARO-LOMBARDO S.R.L.?	False	True
Similar Answer	0.04	Who is the owner of PAVONE SPA?	PERGOLESI, BONOLIS E NONIS GROUP, MILO TORNATORE, FANTINI, DOSSETTI E BROGGINI E FIGLI	PERGOLESI, BONOLIS E NONIS GROUP, GADDA-NATTA E FIGLI
Subset of the Answer	0.03	Who takes a role in DOVARA SPA?	MUTI-DESIO SPA, BRICCIALDI, VIOLA E BORRANI S.R.L.	MUTI-DESIO SPA
Similar Company Type	0.02	Who is the owner of LUNA-SOBRERO S.R.L.?	CHIESA-ZACCO GROUP	SONNINO, GIANINAZZI E ARGENTI GROUP
Wrong Order	0.01	Which company or person has control over RUGGERI-ANTONELLO E FIGLI?	CARACCIOLO-GREGGIO S.R.L., PAGNOTTO, TOMASETTI E VIGORELLI SPA	PAGNOTTO, TOMASETTI E VIGORELLI SPA, CARACCIOLO-GREGGIO S.R.L.

Table 16: Error Analysis for ALIGNed TL-1.1B-Chat on the anonymized CO KG

#	Question	Ground-Truth	P(ALIGNed-TL-1.1B)	P(TL-1.1B)	δ
1	Which company or person assumes a role with SCALFARO S.R.L.?	MUTI-DESIO SPA	1	0.0499	0.9501
2	Which company or person has control over CAGNIN, PAVAROTTI E PIAZZI GROUP?	ABBAGNALE-GRITTI E FIGLI	0.3239	0	0.3239
3	Who takes a role in VOLTERRA, VERDI E BODONI S.R.L.?	CIPOLLA-SANTORO E FIGLI	1	0.3066	0.6934
4	Who takes a role in BABATO-FAVATA SPA?	ORSINI SPA	1	0.284	0.716
5	Who takes a role in GARIBALDI E FIGLI?	CIPOLLA-SANTORO E FIGLI	0.4226	0.2709	0.1517
6	Is it true that FANTINI, DONARELLI E STEFANELLI S.R.L. owns shares of CIPOLLA GROUP?	false	1	0.6654	0.3346

Table 17: Statistics of ALIGNed-LLM answer generation. P(ALIGNed-TL-1.1B) and P(TL-1.1B) is the probability of correct token prediction by ALIGNed-TL-1.1B and fine-tuned version of TL-1.1B respectively. δ denotes the shift in the probability.