Analyzing the factual knowledge of parameter efficient instruction tuned mid-size Large Language Models

Anmol Nayak and Hari Prasad Timmapathini
ARiSE Labs at Bosch
Bangalore, India
{Anmol.Nayak, HariPrasad.Timmapathini}@in.bosch.com

Abstract

Large Language Models (LLM) have significantly improved Natural Language Processing (NLP) by enhancing the accuracy, efficiency, and versatility of various NLP applications, from text generation to language translation, due to their ability to capture and leverage vast amounts of linguistic and factual knowledge. While LLM have pushed the boundaries, they typically need to be further instruction tuned to get improved performance on niche applications. In this paper, we focus on analyzing the factual knowledge of LLM keeping in mind the practical aspects of using LLM by: 1) training only a small injection model (having $\approx$ 0.05 % of the parameters of the base LLM) using the Low Rank Adapation (LoRA) parameter efficient technique, and 2) restricting our study to Llama-2-13b-chat and StableBeluga-13B, which are two mid-size LLM having 13 billion parameters and are based on the LLama 2 architecture. The injection model is instruction tuned for Knowledge Base (KB) construction on the LM-KBC 2023 challenge dataset, which contains subject-relation-object triplets of Wikipedia entities across 21 different factual relations. Our empirical analysis shows that even after instruction tuning, the LLM are: 1) deficient in foundational knowledge of many must-know areas like Geography, 2) unable to effectively use the context supplied in the prompt, and 3) fragile to subtle changes in prompt at inference. The source code for our experiments can be found at: `https://github.com/Ffc1234/NIPS_ICBINB_submission`

1 Introduction and Related Works

The rapid progress in natural language processing (NLP) has driven the creation of large language models, fundamentally transforming how machines comprehend and produce human language. Among the crucial uses of these advanced models is their role in building knowledge repositories, vital for various NLP functions such as information retrieval, question answering, and knowledge inference.

LLM like GPT-4 [1], Llama 2 [2] Stable Beluga 2 [3] excel at comprehending and contextualizing factual information from extensive text sources. KB are typically represented as subject-relation-object triplets for providing structured storage of factual data. Using LLM to build a KB serves as a true test to assess the extent of factual knowledge they possess. However, the effectiveness of building a KB with LLM relies heavily on the capacity to fine-tune these models efficiently. Traditional fine-tuning approaches frequently encounter scalability problems and require substantial computational resources. Nevertheless, recent progress in parameter-efficient fine-tuning strategies, such as LoRA (Low-Rank Adaptation) [4], has exhibited encouraging outcomes by mitigating complexity and computational demands while maintaining model effectiveness. This is because the base model is

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

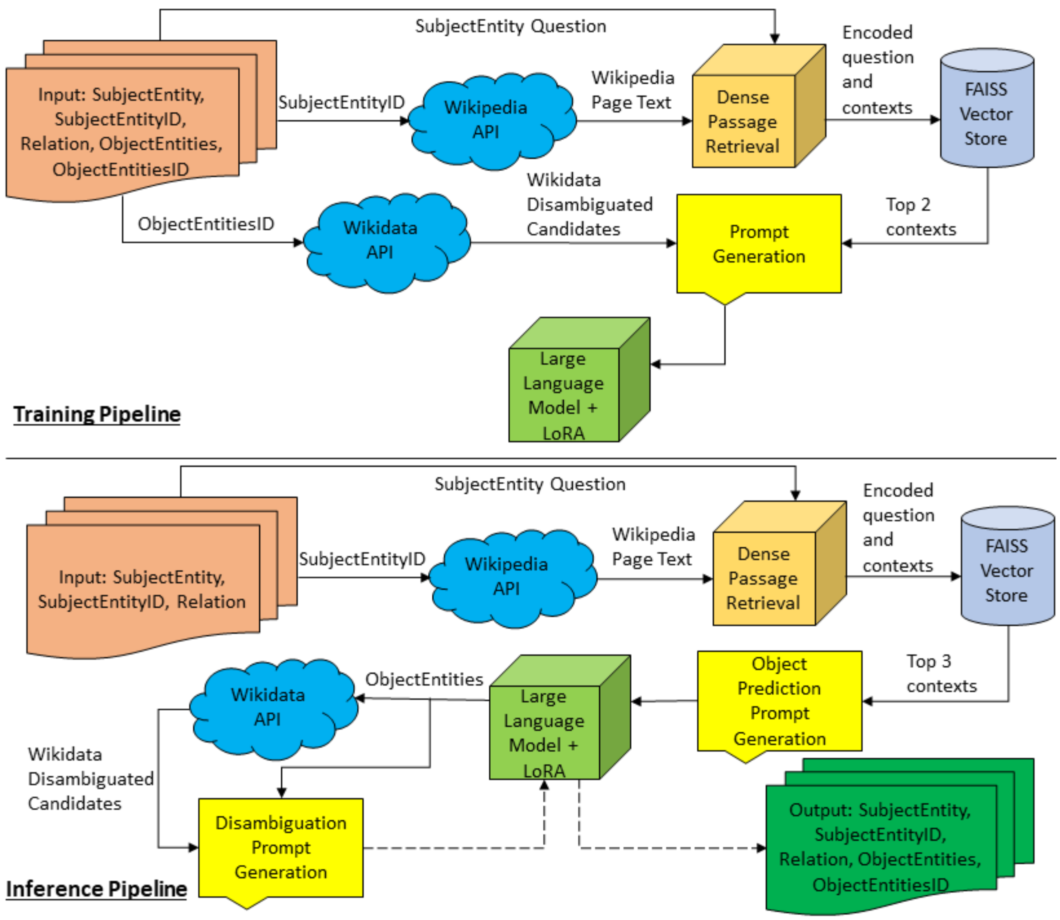

Figure 1: System Architecture.

kept frozen and only a small injection model is trained to handle the incremental instruction tuning required for the application.

There have been many works on extracting factual knowledge through methods like probing language models with masked language modeling to complete cloze-style sentences or various prompt engineering techniques for LLM [5-11], however to the best of our knowledge there is no work which has focused on extracting factual knowledge from LLM when an instruction tuned injection model is plugged along with it. We believe this analysis is required since instruction tuning with small injection models is quickly becoming the norm for fine tuning LLM. The task we instruction tune the injection model to solve is that given a subject-entity and a relation, predict all correct object-entities (o1, o2, ..., ok).

2 Experiment Setup

2.1 Models

We have chosen the 'Llama-2-13b-chat' model from Meta AI and the 'StableBeluga-13B' model from Stability AI as our foundational models because they have undergone fine-tuning for instruction following, are openly available, and have demonstrated top-tier performance across various benchmarks. For the retrieval of subject entity context, we used 'dpr-ctx_encoder-single-nq-base' for encoding questions and 'dpr-ctx_encoder-multiset-base' for encoding contexts, which are Dense Passage Retrieval (DPR) models from Meta AI, as detailed in [12]. DPR encompasses a suite of tools and models widely utilized in cutting-edge open-domain Question and Answer research. The training compute and hyper-parameter details can be found in Appendix A.

Table 1: Questions corresponding to each relation. _ is the placeholder for the Subject Entity.

Relation	**Question**
BandHasMember	Who are the members of _?
CityLocatedAtRiver	Which river is _ located at?
CompanyHasParentOrganisation	What is the parent organization of _?
CompoundHasParts	What are the components of _?
CountryBordersCountry	Which countries border _?
CountryHasOfficialLanguage	What is the official language of _?
CountryHasStates	Which states are part of _?
FootballerPlaysPosition	What position does _ play in football?
PersonCauseOfDeath	What caused the death of _?
PersonHasAutobiography	What is the title of _'s autobiography?
PersonHasEmployer	Who is _'s employer?
PersonHasNoblePrize	In which field did _ receive the Nobel Prize?
PersonHasNumberOfChildren	How many children does _ have?
PersonHasPlaceOfDeath	Where did _ die?
PersonHasProfession	What is _'s profession?
PersonHasSpouse	Who is _ married to?
PersonPlaysInstrument	What instrument does _ play?
PersonSpeaksLanguage	What languages does _ speak?
RiverBasinsCountry	In which country can you find the _ river basin?
SeriesHasNumberOfEpisodes	How many episodes does the series _ have?
StateBordersState	Which states border the state of _?

2.2 Dataset creation

The LM-KBC dataset [13] comprises of train, validation and test set, containing 1940 records each and across 21 diverse relations. Each records includes the subject-entity Wikidata ID, subject-entity name, a list of all possible object-entity Wikidata IDs, a list of all possible object-entities, and the corresponding relation. We transform this dataset into a format suitable for instruction tuning in the following way: First, the Train set and Validation set are combined to produce a super set. With each record, we begin generating a separate instruction tuning dataset (see Section 2.3 for details on the Prompts) by slot filling Prompt 1, 2 and 3. Prompt 2 is the same as Prompt 1 but with an empty context. Prompt 3 performs entity disambiguation.

Thus, each record in the super set begins by producing 2 samples using Prompt 1 and 2 for the instruction tuning dataset. Further, for each ObjectEntity in the record we slot fill Prompt 3, which additionally generates as many new samples for the instruction tuning dataset as there are object entities (since each valid ObjectEntity is posed as a separate disambiguation task to the LLM). The generated instruction tuning dataset is shuffled, keeping 14310 samples for training and 1000 samples for validation of the model. The test set is retained as it is to evaluate the performance of the system.

2.3 LLM Prompts

We created 4 different prompts based on the format of each of the base models, where Prompts 1,2 and 3 are used for performing instruction tuning and Prompt 4 is used only during inference for in-context learning (formatting the output into a valid Python list). Note: During inference, Prompts 1,2 and 3 are trimmed after *[/INST]* for Llama-2-13b-chat and after *Answer:* for StableBeluga-13B. Instruction tuning is a technique to perform supervised fine tuning of models to make them learn to produce valid responses for specific instructions. Prompt 1,2 and 3 are expecting the LLM to produce an answer in a Python list of string format, however we noticed that the model sometimes produces a syntactically incorrect list. Hence, we use Prompt 4 to demonstrate to the LLM an incorrect format Answer vs. correct format Answer. Prompt 1 is as follows:

<s>[INST] «SYS»

You are a helpful, respectful and honest assistant. Your answers should be crisp, short and not repititive.

Give valid wikipedia page titles in the answer. The answer should be in a python list of string format.

If you dont know the answer from both the given context and your past knowledge, answer should just be a python empty list.

«/SYS»

context: '{context}'

{question} [/INST] Answer: {answer} </s>

In this prompt, the {question} variable is formed by first picking the relevant question corresponding to the relation (see Table 1) and then replacing the subject entity of the data sample in the {question}. {context} variable is formed by concatenating the strings of the top 2 contexts returned by the DPR system for the given {question}. It is important to note that to fetch the top 2 contexts, the DPR context encoder is only fed the Wikipedia page textual content of the subject entity and we do not use the content of the Wikipedia Infobox (found at the top right corner of a Wikipedia page), since the Infobox already contains semi-structured information about an entity which would defeat the purpose of using LLM. The {answer} variable is formed by replacing the object entities of the data sample. Due to space constraints, we explain the remaining prompts for Llama and Beluga in Appendix B and Appendix C respectively.

2.4 Inference

The overall architecture can be seen in Fig. 1. During inference, we first try to fetch the English language Wikipedia page text of the subject entity using its Wikidata ID. If the subject entity does not have an English page, we then pick the text from the primary alternate language page. We then split the text into chunks of 300 tokens (with an overlap of 50 tokens to maintain continuity). Each of the context chunks are encoded using the DPR context encoder and stored in a Facebook AI Similarity Search (FAISS) vector store for fast search and retrieval.

To ensure that the LLM is not just fixated on using only top 2 contexts, during inference we pick top 3 contexts to test its robustness towards handling variability. To pick the top 3 relevant context chunks for a given question, the question is encoded using the DPR question encoder and then passed to FAISS. The top 3 retrieved context chunks are concatenated to form the {context} variable for the prompts. In cases where a subject entity does not have a Wikipedia page, the {context} variable will be empty for the prompt, and the LLM will have to rely on their stored knowledge to answer the question. Prompt 1 is executed if a context was found whereas Prompt 2 is executed if no context was found.

Once the LLM processes the prompt, the answer can have 0 or more object entities. For disambiguating each of these surface strings, we query the Wikidata API with each of these object entities separately. The API will return a list of candidate Wikidata entities, from which we attempt to find the correct disambiguated entity. To do this, we collect all the candidate entities and put them in a list to form the context in Prompt 3. The expectation is that the LLM picks the correct disambiguated entity relevant to the question. The Wikidata ID of each disambiguated object entity is then fetched and stored. In cases where the LLM generated a string which was not an exact match of the provided options, we pass the LLM output to Wikipedia API for fetching the most relevant entity.

3 Results and Conclusion

The performance of our system with the LLM+LoRA injection model can be seen in Table 2 for Llama-2-13b-chat and in the Table 3 for StableBeluga-13B. Across the 21 factual relations, the average F1 scores obtained by Llama-2-13b-chat and StableBeluga-13B were 0.6185 and 0.5520 respectively. These are significant improvements over the baseline LLM scores of 0.218 (Few-shot GPT-3 for object prediction + Wikidata API for entity disambiguation) and 0.061 (Few-shot GPT-3 for directly predicting the object entities Wikidata ID) reported by the LM-KBC 2023 challenge, with our best performing Llama-2-13b-chat model being ranked #4 overall and ranked #1 in precision (beating even GPT-4 based approaches), we believe that training small plug-and-play injection models can improve precision of factual knowledge retrieved from mid-size LLM making them competitive with large-size LLM, however the recall is still largely correlated to the size of LLM (due to its

Table 2: Llama-2-13b-chat Test set results (Precision, Recall and F1 score are macro average per relation).

Relation	Precision	Recall	F1 score
BandHasMember	0.7178	0.4151	0.4857
CityLocatedAtRiver	0.7400	0.5205	0.5586
CompanyHasParentOrganisation	0.7200	0.7200	0.6400
CompoundHasParts	0.9318	0.8495	0.8784
CountryBordersCountry	0.8492	0.5848	0.6719
CountryHasOfficialLanguage	0.9154	0.7500	0.8015
CountryHasStates	0.5687	0.3872	0.4399
FootballerPlaysPosition	0.7050	0.6683	0.6783
PersonCauseOfDeath	0.8500	0.8200	0.8200
PersonHasAutobiography	0.6700	0.4075	0.4173
PersonHasEmployer	0.5100	0.2957	0.3327
PersonHasNoblePrize	0.9900	0.9350	0.9367
PersonHasNumberOfChildren	0.4900	0.4900	0.4900
PersonHasPlaceOfDeath	0.8300	0.7600	0.7200
PersonHasProfession	0.6600	0.4628	0.5034
PersonHasSpouse	0.7700	0.5850	0.5867
PersonPlaysInstrument	0.7633	0.6159	0.6354
PersonSpeaksLanguage	0.8500	0.7538	0.7704
RiverBasinsCountry	0.8883	0.7689	0.7896
SeriesHasNumberOfEpisodes	0.4500	0.4500	0.4500
StateBordersState	0.5090	0.3416	0.3815
Average	**0.7323**	**0.5991**	**0.6185**

Table 3: StableBeluga-13B Test set results (Precision, Recall and F1 score are macro average per relation).

Relation	Precision	Recall	F1 score
BandHasMember	0.7075	0.4227	0.4850
CityLocatedAtRiver	0.7300	0.4865	0.5225
CompanyHasParentOrganisation	0.8800	0.6500	0.6300
CompoundHasParts	0.8573	0.6581	0.7197
CountryBordersCountry	0.8360	0.4690	0.5654
CountryHasOfficialLanguage	0.9462	0.7372	0.8021
CountryHasStates	0.7984	0.3225	0.3853
FootballerPlaysPosition	0.7100	0.4933	0.5050
PersonCauseOfDeath	0.8800	0.8000	0.8000
PersonHasAutobiography	0.8100	0.3450	0.3533
PersonHasEmployer	0.6700	0.2128	0.2403
PersonHasNoblePrize	0.9800	0.7700	0.7700
PersonHasNumberOfChildren	0.4400	0.4400	0.4400
PersonHasPlaceOfDeath	0.9400	0.6700	0.6700
PersonHasProfession	0.6100	0.3962	0.4457
PersonHasSpouse	0.8500	0.4450	0.4467
PersonPlaysInstrument	0.7667	0.5577	0.6055
PersonSpeaksLanguage	0.8850	0.6190	0.6947
RiverBasinsCountry	0.9170	0.7684	0.8030
SeriesHasNumberOfEpisodes	0.4200	0.4200	0.4200
StateBordersState	0.6817	0.2269	0.2872
Average	**0.7769**	**0.5195**	**0.5520**

ability to store more knowledge). Some of the practical challenges in using LLM for Knowledge Base Construction that we found during our experiments are described in the following sub-sections. Note: Due to space constraints we only provide 1 example per challenge, however this behaviour happens frequently across the test set and can be analyzed by viewing our GitHub repository.

3.1 Deficient foundational knowledge on fundamental topics

Among the 21 relations, we believe that the 6 Geographical relations (CityLocatedAtRiver, CountryBordersCountry, CountryHasOfficialLanguage, CountryHasStates, StateBordersState, RiverBasinsCountry) and the 1 Scientific relation (CompoundHasParts) will have Object entities that rarely change and can be considered to be fundamental must-know facts for the LLM. The average F1 score across these 7 relations was 0.6459 for Llama-2-13b-chat and 0.584 for StableBeluga-13B. Even after refreshing the factual memory of the LLM with instruction tuning, they do not perform significantly well for these must-know facts even though the pre-training data for these LLM contain Wikipedia information. We believe this behaviour can be attributed to the fact although we have done instruction tuning, it does not inject any new factual knowledge but only helps in refreshing the LLM memory and realigning the output to a certain format.

For e.g. when Llama-2-13b-chat was prompted with the question *Which countries border India?*, it fails to answer China (even though it is part of the context), Bangladesh and Bhutan from its memory (please refer to Appendix Section D.1 for the complete prompt and LLM output).

The other 14 relations, although factual in nature, can have Object entities that change over time. We believe this could have impacted the scores (to some extent although not significantly) since every model has knowledge only until a cut-off period. For e.g. a TV Series could have 8 episodes at the time the LLM was trained but later gets another 10 episodes.

3.2 Ineffective usage of context

The effective usage of the supplied Wikipedia context can significantly boost the LLM performance on both types of factual relations but is especially critical for handling the 14 temporal factual relations since it can help the LLM to overrule its outdated knowledge. The average F1 score across these 14 relations was 0.6047 for Llama-2-13b-chat and 0.5361 for StableBeluga-13B. Even though the LLM were supplied with the top 3 relevant contexts from the subject entities Wikipedia page, the LLM failed to utilize it effectively. This behaviour was described in a recent work [14], which showed that LLM are weak in grounding their output based on relevant context (controllability) and ignoring irrelevant context (robustness). We hoped that instruction tuning the LLM injection model with (Prompt 1) and without context (Prompt 2) samples would alleviate this problem and result in better controllability and robustness, however this did not happen consistently.

For e.g. when Llama-2-13b-chat was prompted with the question *How many children does Anne Igartiburu have?*, it fails to answer 3 even though the three children were mentioned in the context (please refer to Appendix Section D.2 for the complete prompt and LLM output).

3.3 Fragility in handling subtle changes to prompts

While prompt brittleness and hallucinations are well known drawbacks of existing LLM and is an ongoing field of research, our work is specifically focused on analyzing whether these effects are alleviated by instruction tuning an injection model that works in tandem with the base LLM. The expectation is that since instruction tuning helps the LLM adapt to the user's objective of adhering to human instructions, it is naturally favourable towards reducing the chance of the LLM being fragile to subtle changes in prompt.

We found that while instruction tuning significantly reduced the hallucinations in comparison to just using the base LLM without the injection model, this held true only as long as the prompts used at inference were exactly the same as in training. Even slight deviations from the prompt template made the LLM to begin rambling.

For e.g. when Llama-2-13b-chat was prompted with the question *What is the official language of Greece?*, it gives the right answer only if the [/INST] token is immediately following the question and fails when a simple change like adding a new line occurs. (please refer to Appendix Section D.3 for the complete prompt and LLM output).

References

[1] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[3] Stability ai stable beluga, 2023. URL: https://stability.ai/blog/stable-beluga-large-instruction-fine-tuned-models.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[5] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. Miller, Language models as knowledge bases?, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 2463–2473.

[6] S. Razniewskia, A. Yatesa, N. Kassnerc, G. Weikuma, Language models as or for knowledge bases (2021).

[7] Z. Jiang, F. F. Xu, J. Araki, G. Neubig, How can we know what language models know?, Transactions of the Association for Computational Linguistics 8 (2020) 423–438.

[8] Z. Bouraoui, J. Camacho-Collados, S. Schockaert, Inducing relational knowledge from bert, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 7456–7463.

[9] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, S. Singh, Autoprompt: Eliciting knowledge from language models with automatically generated prompts, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4222–4235.

[10] T. Li, W. Huang, N. Papasarantopoulos, P. Vougiouklis, J. Z. Pan, Task-specific pre-training and prompt decomposition for knowledge graph population with language models, CEUR- WS (2022).

[11] D. Alivanistos, S. B. Santamaría, M. Cochez, J.-C. Kalo, E. van Krieken, T. Thanapalasingam, Prompting as probing: Using language models for knowledge base construction, CEUR-WS (2022).

[12] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781.

[13] S. Singhania, J.-C. Kalo, S. Razniewski, J. Z. Pan, Lm-kbc: Knowledge base construction from pre-trainedlanguage models, semantic web challenge @ iswc, CEUR-WS (2023). URL: https://lm-kbc.github.io/challenge2023/.

[14] D. Li, A. S. Rawat, M. Zaheer, X. Wang, M. Lukasik, A. Veit, F. Yu, S. Kumar. Large language models with controllable working memory. arXiv preprint arXiv:2211.05110 (2022).

[15] T. Dettmers, M. Lewis, Y. Belkada, L. Zettlemoyer, Llm. int8 (): 8-bit matrix multiplication for transformers at scale, arXiv preprint arXiv:2208.07339 (2022).

[16] T. Dettmers, M. Lewis, S. Shleifer, L. Zettlemoyer, 8-bit optimizers via block-wise quantization, 9th International Conference on Learning Representations, ICLR (2022).

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp- demos.6.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019).

A Training hyper-parameter settings

For both base LLM namely Llama-2-13b-chat and StableBeluga-13B, we load them in a 4 bit quantized state with frozen weights and only train an injection model using LoRA technique. The injection model has ≈ 0.05 % of the parameters of the base models. The time taken for training was ≈ 9 hours.

- **Libraries**: BitsandBytes [15, 16], HuggingFace [17], PyTorch [18]
- **Base model**: BitsandBytesConfig(load_in_4bit=True, bnb_4bit_use_double_quant=True, bnb_4bit_quant_type="nf4", bnb_4bit_compute_dtype=torch.bfloat16)
- **LoRA model**: alpha=16, dropout=0.05, r=4, bias="none", task_type="CAUSAL_LM"
- **Trainer**: epochs=3, optimizer="paged_adamw_32bit", gradient_accumulation_steps=2, per_device_train_batch_size=1, fp16=True, learning_rate=2e-5, max_grad_norm=0.3, warmup_ratio=0.03, lr_scheduler_type="constant", evaluation_strategy="epoch", per_device_eval_batch_size= 4
- **GPU**: 2x NVIDIA V100

B Llama-2-13b-chat Prompts

- Prompt 2:

  <s>[INST] «SYS»

  You are a helpful, respectful and honest assistant. Your answers should be crisp, short and not repititive.

  Give valid wikipedia page titles in the answer. The answer should be in a python list of string format.

  If you dont know the answer from both the given context and your past knowledge, answer should just be a python empty list.

  «/SYS»

  context: ”

  {question} [/INST] Answer: {answer} </s>

  In this prompt, the {question} variable and {answer} variable are formed in the same method as prompt 1 as described in Section 2.3, however we do not supply any context.

- Prompt 3:

  <s>[INST] «SYS»

  You are a helpful, respectful and honest assistant. Your answers should be crisp, short and not repititive.

  Choose an answer from the options in the context.

  If you dont know the answer from the given context, answer should just be a python empty list.

  «/SYS»

  context: '{options}'

  {question} [/INST] Answer: {answer} </s>

  In this prompt, the {question} variable is formed in the same method as the previous prompts, the {options} variable is a list of titles (appended with their descriptions) for the Wikidata entities returned by the API when queried with a given object entity text. The titles and descriptions are fetched from the API response. Each separate object entity will lead to 1 unique training sample. For e.g. if a data sample has 4 object entities, then we will generate 4 separate training samples for it. The {answer} variable is the title of the response entity which has the same Wikidata ID as the object entity.

- Prompt 4:

  <s>[INST] «SYS»

Example 1: Wrong Format: ['People's Republic of China', 'Laos', 'Thailand', 'India', 'Bangladesh']"]. Correct Format: Answer: ["People's Republic of China", "Laos", "Thailand", "India", "Bangladesh"] </s>

Example 2: Wrong Format: ['Artibonite', 'Nord-Est Department', 'South Department','West Department', 'Centre Department', 'Grand'Anse Department', 'North Department']. Correct Format: Answer: ["Artibonite", "Nord-Est Department", "South Department", "West Department", "Centre Department", "Grand'Anse Department", "North Department"] </s>

Example 3: Wrong Format: ['book's and page's']. Correct Format: Answer: ["book's and page's"] </s>

Your answer should only be a valid python list of string format. Do not give any explainations.

«/SYS»

Use the examples to convert {answer} into a correct python list. [/INST] Answer:

Once an answer is generated from Prompt 1, 2 or 3, it is used as the {answer} variable in this prompt to format the answer in a correct Python list of string format.

C StableBeluga-13B Prompts

- Prompt 1:

  ### System:

  You are a helpful, respectful and honest assistant. Your answers should be crisp, short and not repititive.

  Give valid wikipedia page titles in the answer. The answer should be in a python list of string format.

  If you dont know the answer from both the given context and your past knowledge, answer should just be a python empty list.

  ### User:

  context: {context}

  {question}

  ### Assistant

  Answer: {answer}

- Prompt 2:

  ### System:

  You are a helpful, respectful and honest assistant. Your answers should be crisp, short and not repititive.

  Give valid wikipedia page titles in the answer. The answer should be in a python list of string format.

  If you dont know the answer from both the given context and your past knowledge, answer should just be a python empty list.

  ### User:

  context: "

  {question}

  ### Assistant

  Answer: {answer}

- Prompt 3:

  ### System:

  You are a helpful, respectful and honest assistant. Your answers should be crisp, short and not repititive.

  Choose an answer from the options in the context.

  If you dont know the answer from both the given context, answer should just be a python empty list.

User:
context: {options}
{question}
Assistant
Answer: ['{answer}']

- Prompt 4:
  ### System:
  Example 1: Wrong Format: ['People's Republic of China', 'Laos', 'Thailand', 'India', 'Bangladesh']"]. Correct Format: Answer: ["People's Republic of China", "Laos", "Thailand", "India", "Bangladesh"] </s>
  Example 2: Wrong Format: ['Artibonite', 'Nord-Est Department', 'South Department', 'West Department', 'Centre Department', 'Grand'Anse Department', 'North Department'].
  Correct Format: Answer: ["Artibonite", "Nord-Est Department", "South Department", "West Department", "Centre Department", "Grand'Anse Department", "North Department"] </s>
  Example 3: Wrong Format: ['book's and page's']. Correct Format: Answer: ["book's and page's"] </s>
  ### User:
  Your answer should only be a valid python list of string format. Do not give any explainations.
  Use the examples to convert {answer} into a correct python list.
  ### Assistant
  Answer:

D Prompts showing the challenges of LLM

D.1 Deficient foundational knowledge on fundamental topics

<s> [INST] «SYS»

You are a helpful, respectful and honest assistant. Your answers should be crisp, short and not repititive.

Give valid wikipedia page titles in the answer. The answer should be in a python list of string format.

If you dont know the answer from both the given context and your past knowledge, answer should just be a python empty list.

«/SYS»

context: 'partitioned into two independent dominions, a Hindu-majority Dominion of India and a Muslim-majority Dominion of Pakistan, amid large-scale loss of life and an unprecedented migration.India has been a federal republic since 1950, governed through a democratic parliamentary system. It is a pluralistic, multilingual and multi-ethnic society. India's population grew from 361 million in 1951 to almost 1.4 billion in 2022. During the same time, its nominal per capita income increased from US$64 annually to US$2,601, and its literacy rate from 16.6% to 74%. From being a comparatively destitute country in 1951, India has become a fast-growing major economy and a hub for information technology services, with an expanding middle class. It has a space programme. Indian movies, music, and spiritual teachings play an increasing role in global culture. India has substantially reduced its rate of poverty, though at the cost of increasing economic inequality. India is a nuclear-weapon state, which ranks high in military expenditure. It has disputes over Kashmir with its neighbours, Pakistan and China, unresolved since the mid-20th century. Among the socio-economic challenges India faces are gender inequality, child malnutrition, and rising levels of air pollution. India's land is megadiverse, with four biodiversity hotspots. Its forest cover comprises 21.7% of its area. India's wildlife, which has traditionally been viewed with tolerance in India's culture, is supported among these forests, and elsewhere, in protected habitats.

unique among the world's newer nations; however, in spite of its recent economic successes, freedom from want for its disadvantaged population remains a goal yet to be achieved.

China's nuclear test of 1964, as well as its repeated threats to intervene in support of Pakistan in the 1965 war, convinced India to develop nuclear weapons. India conducted its first nuclear weapons test in 1974 and carried out additional underground testing in 1998. Despite criticism and military sanctions, India has signed neither the Comprehensive Nuclear-Test-Ban Treaty nor the Nuclear Non-Proliferation Treaty, considering both to be flawed and discriminatory. India maintains a "no first use" nuclear policy and is developing a nuclear triad capability as a part of its "Minimum Credible Deterrence" doctrine. It is developing a ballistic missile defence shield and, a fifth-generation fighter jet. Other indigenous military projects involve the design and implementation of Vikrant-class aircraft carriers and Arihant-class nuclear submarines.Since the end of the Cold War, India has increased its economic, strategic, and military co-operation with the United States and the European Union. In 2008, a civilian nuclear agreement was signed between India and the United States. Although India possessed nuclear weapons at the time and was not a party to the Nuclear Non-Proliferation Treaty, it received waivers from the International Atomic Energy Agency and the Nuclear Suppliers Group, ending earlier restrictions on India's nuclear technology and commerce. As a consequence, India became the sixth de facto nuclear weapons state. India subsequently signed co-operation agreements involving civilian nuclear energy with Russia, France, the United Kingdom, and Canada.'

Which countries border India? [/INST] Answer: ['Nepal', 'Pakistan', 'Myanmar', 'Afghanistan'] </s>

D.2 Ineffective usage of context

<s> [INST] «SYS»

You are a helpful, respectful and honest assistant. Your answers should be crisp, short and not repititive.

Give valid wikipedia page titles in the answer. The answer should be in a python list of string format.

If you dont know the answer from both the given context and your past knowledge, answer should just be a python empty list.

«/SYS»

context: 'Personal life She was married to ballet dancer Igor Yebra from 2004 to 2006. Igartiburu has two adopted daughters, Noa (born 2000 in India) and Carmen (born 2011 in Vietnam). In 2015, she married conductor Pablo Heras-Casado. On 13 June 2016 she gave birth to their first baby son, Nicolás.

Biography She studied Industrial Marketing. She started working in the local television of Mondragón in 1993. She then worked in Euskal Telebista, and later in Telecinco, hosting Una pareja feliz alongside Antonio Hidalgo in the 1994–95 season. Since 1997, Igartiburu has presented daily celebrity news programme Corazón on Televisión Española, except for a period in 2012–13 when she presented daily infotainment show +Gente. From 2005 to 2009 she hosted the Spanish adaptation of Dancing with the Stars (¡Mira quién baila!). In 2006 she hosted the special show Gala 50 años de TVE, which celebrated the 50th anniversary of TVE, alongside Laura Valenzuela and Paula Vázquez.Since 2005, Igartiburu has presented the annual New Year's Eve celebration broadcast for TVE live from Madrid's Puerta del Sol. Igartiburu has also been associated with the Eurovision Song Contest: she was the Spanish spokesperson in the 2002, 2003 and 2004 contests, and she hosted the Spanish national final in 2010, 2011, 2012, 2014, and 2016.

Anne Igartiburu at IMDb'

How many children does Anne Igartiburu have? [/INST] Answer: ['2'] </s>

D.3 Fragility in handling subtle changes to prompts

<s> [INST] «SYS»

You are a helpful, respectful and honest assistant. Your answers should be crisp, short and not repititive.

Give valid wikipedia page titles in the answer. The answer should be in a python list of string format.

If you dont know the answer from both the given context and your past knowledge, answer should just be a python empty list.

«/SYS»

context: 'Greek. The Greek Orthodox Church, which emerged in the first century AD, helped shape modern Greek identity and transmitted Greek traditions to the wider Orthodox world. After falling under Ottoman rule in the mid-15th century, Greece emerged as a modern nation state in 1830 following a war of independence. Over the first hundred years the kingdom of Greece sought its territorial expansion, which was mainly achieved in the early 20th century, during the Balkan Wars and up until its Asia Minor Campaign which ended with a catastrophic defeat in 1922. The short-lived republic that followed, beset by the ramifications of civil strife and the challenge of resettling the refugees from Turkey, came to an end in 1936, when the imposition of a royalist dictatorship inaugurated a long period of authoritarian rule, marked by military occupation during World War II, civil war and military dictatorship. Greece achieved record economic growth from 1950 through the 1970s, allowing it to join the ranks of developed nations. Democracy was restored in 1974–75, and Greece has been a parliamentary republic ever since. The country's rich historical legacy is reflected in part by its 18 UNESCO World Heritage Sites. Greece is a unitary parliamentary republic, and a developed country, with an advanced high-income economy. Its economy is the second largest in the Balkans, where it is an important regional investor. A founding member of the United Nations, Greece was the tenth member to join the European Communities (precursor to the European Union) and has been part of the Eurozone since 2001. It is also a member of numerous other international institutions, including the Council of Europe, NATO, the OECD, the WTO, and the OSCE. Greece has a unique cultural heritage, large tourism industry, and prominent shipping sector.

Greek language and Greek culture in the territories conquered by Alexander. Greek science, technology, and mathematics are generally considered to have reached their peak during the Hellenistic period.

Languages Greece is today relatively homogeneous in linguistic terms, with a large majority of the native population using Greek as their first or only language. Among the Greek-speaking population, speakers of the distinctive Pontic dialect came to Greece from Asia Minor after the Greek genocide and constitute a sizable group. The Cappadocian dialect came to Greece due to the genocide as well, but is endangered and is barely spoken now. Indigenous Greek dialects include the archaic Greek spoken by the Sarakatsani, traditionally transhument mountain shepherds of Greek Macedonia and other parts of Northern Greece. The Tsakonian language, a distinct Greek language deriving from Doric Greek instead of Koine Greek, is still spoken in some villages in the southeastern Peloponnese. The Muslim minority in Thrace, which amounts to approximately 0.95% of the total population, consists of speakers of Turkish, Bulgarian (Pomaks) and Romani. Romani is also spoken by Christian Roma in other parts of the country. Further minority languages have traditionally been spoken by regional population groups in various parts of the country. Their use has decreased radically in the course of the 20th century through assimilation with the Greek-speaking majority. Today they are only maintained by the older generations and are on the verge of extinction. The same goes for the Arvanites, an Albanian-speaking group mostly located in the rural areas around the capital Athens, and for the Aromanians and Megleno-Romanians, also known as "Vlachs", whose language is closely related to Romanian and who used to live scattered across several areas of mountainous central Greece. Members of these groups usually identify ethnically as Greek and are today all at least bilingual in Greek. Near the northern Greek borders there are also some Slavic–speaking groups, locally known as Slavomacedonian-speaking, most of whose members identify ethnically as Greeks. It is estimated that'

- **Question and incorrect predicted Answer with [/INST] on a new line:**

  What is the official language of Greece?

  [/INST]

  Language(s): Greek - Greek is the official language of Greece. It belongs to the family of Indo-European Languages. It is spoken by 9,900,000 people within the country. There are 1,300,000 Greek speakers in Cyprus, 300,000 in Albania, 200,000 in Former Yugoslav Republic of Macedonia, 100,000 in Bulgaria, 100,000 in Turkey, 50,000 in Italy, 20,000 in Australia, 10,000 in

- **Question and correct predicted Answer with [/INST] on the same line:**

What is the official language of Greece? [/INST]
Answer: ['Greek'] </s>