Expanding Horizons or Hitting Walls? Limits and Potentials of LLMs in Augmenting Lexical Knowledge Bases

Anonymous ACL submission

Abstract

This paper investigates the potential of Large Language Models (LLMs) to augment lexical knowledge bases (KBs) and to address their common limitations, such as static nature, limited coverage, and labor-intensive creation and maintenance. We propose a methodology that leverages LLMs to accurately reconstruct information from a source KB and generate new knowledge. Then, we evaluate this methodology using various LLMs and prompting techniques across three separate KBs. The results suggest that LLMs can accurately provide information when given ample contextual cues and when dealing with high-specificity concepts. However, they are prone to errors and inconsistencies when asked for rare or generic knowledge. The findings also indicate that LLMs can contribute to KB management by reducing the need for manual intervention. This study highlights the potential and limitations of LLMs in lexical semantics and emphasizes the importance of novel approaches to KB creation, maintenance, and integration.

1 Introduction

Lexical semantics represents a foundational aspect of Natural Language Processing (NLP), serving as the intersection where the meanings of words and their interrelationships converge. This discipline has always seen unstructured data become structured through the means of knowledge bases (KBs). These latter ones, however, face three common limitations: *i*) they exhibit a static nature, making it challenging to adapt to domains evolution; *ii*) they suffer from limited coverage, hindering their applicability across diverse domains; *iii*) their creation and maintenance is typically laborious, involving human-in-the-loop procedures.

The rise of Large Language Models (LLMs) within Generative AI highlights the importance of interpretable knowledge encapsulation, with KBs being crucial for both enhancing LLM training and providing a means of error, inconsistency, and bias checking (Pan et al., 2024). This necessity becomes particularly pronounced given the expanding influence of Generative AI and its accompanying challenges, including issues such as hallucination (Ji et al., 2023).

This paper unveils an innovative methodology grounded in LLMs to tackle pivotal concerns within lexical semantics. In particular, our contribution is three-fold: *i*) we harness LLMs to reconstitute information encapsulated in a source KB to test their proficiency on this task; *ii*) subsequently, LLMs are deployed to create novel knowledge, proving their aptitude in crafting, and encoding KBs; *iii*) through a third-phase assessment of the newly generated content, we can finally evaluate the capability to expand upon the original KBs and, consequently, assess their completeness.

By conducting thorough experiments utilizing diverse LLMs and prompting techniques across three separate KBs, we elucidate the capacity of LLMs to furnish accurate information, particularly when supplied with substantial contextual cues and when dealing with concepts of high specificity. When confronted with requests for rare or generic knowledge, LLMs are instead prone to errors and inconsistencies.

2 Related Work

In the context of this work, it is essential to clarify that lexical semantic resources, KBs, ontologies and knowledge graphs represent facets or interpretations of the same underlying subject matter.

2.1 Knowledge Acquisition: KBs and KGs

Construction of KBs involves both manual and automatic methods, with famous KBs like WordNet (Fellbaum, 2020) and ConceptNet (Speer et al., 2017) initially depending on manual input. To reduce labor, automated IE techniques have been developed (Fader et al., 2011; Angeli et al., 2015;

Vo and Bagheri, 2017), extracting information from texts to update KBs. ML and NLP progress has also advanced in automatic Knowledge Graph (KG) construction, utilizing data to enhance traditional approaches (Chen et al., 2021).

2.2 Large Language Models

Recently, the advent of LLMs has opened new avenues for knowledge acquisition and representation. LLMs, such as GPT-3 (Brown et al., 2020) and LLama-2 (Touvron et al., 2023), have demonstrated remarkable capabilities in understanding and generating natural language text. Researchers have begun exploring the knowledge encoded within LLMs, probing their ability to serve as implicit KBs (Petroni et al., 2019; Razniewski and Weikum, 2021). This approach offers a novel means of accessing vast amounts of knowledge without explicit curation, although challenges remain in interpreting and validating the knowledge encoded in these models (Chang et al., 2023).

Despite the speed of breakthrough advancements in the field, LLMs still grapple with issues that fall into two main categories: architectural and data-related problems. Architectural problems are inherent to the model's structure and necessitate a change in architecture for resolution. These include the prompt engineering problem, wherein models are non-deterministic and require the "perfect" prompt to elicit the correct response, as highlighted by Park et al. (2022). Conversely, data-related problems stem from the training methodologies and the datasets used, affecting the models' mathematical and reasoning capabilities (Imani et al., 2023; Hendrycks et al., 2021), as well as their common sense understanding (West et al., 2022).

2.3 KBs/KGs and LLMs

Petersen and Potts (2023) demonstrate LLMs' capability to interpret the word "break" and suggest that these models can advance lexical semantics. Their analysis reveals LLMs' proficiency in identifying both known and novel meanings, as well as their superiority in semantic analysis. Kandpal et al. (2023) indicate that the knowledge representation in LLM training data affects content generation accuracy, particularly for uncommon concepts, challenging our understanding of LLMs' semantic encoding. Cohen et al. (2023) propose a new approach for examining LLM knowledge using graph-based queries, which aligns with our emphasis on structured prompts to retrieve and leverage knowledge from LLMs. In contrast to previous efforts, our work presents a scalable pipeline for standardized KB extension, leveraging LLMs and human-in-the-loop evaluation techniques.

3 A Methodology for KB Extension

Our proposed methodology encompasses different key modules, answering the following two main research questions: *RQ1) How well can LLMs mimic KB concepts and relationships?*; and *RQ2) Do LLMs possess the capability to produce novel information suitable for integration into existing KBs?* However, these inquiries serve as gateways to further exploration. Particularly in relation to *RQ1: what factors of both LLMs and KBs impact the quality of generated content?* This paper delves into the following considerations: LLM architecture (pre-trained, fine-tuned, and storytelling-oriented); prompting and extraction techniques (zero-shot versus one-shot); as well as the scale and intricacy of the KBs. Additionally, in relation to *RQ2: how does the quality of newly generated content compare to that of the original KB?* A manual assessment could provide insights into the completeness of the original resource, enhancing the proposed framework.

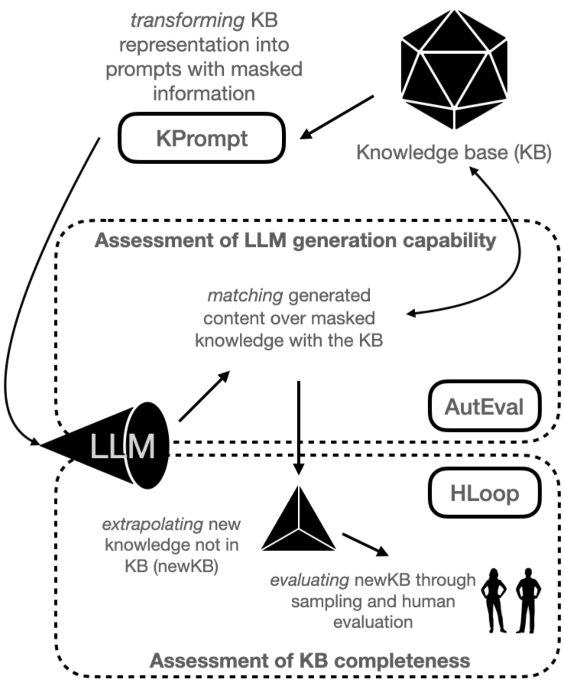

Figure 1: Architecture of the proposed framework for KBs extension. *i)* (*KPrompt*) encodes the source knowledge into masked prompts, that LLMs use for *ii)* (*AutEval*) re-generating existing knowledge and *iii)* (*HLoop*) generate new content to be manually-validated.

On this direction, the proposed framework includes three modules which are designed *i)* to systematically assess the proficiency of LLMs in delivering concepts aligned with existing KBs, *ii)* to probe their ability to generate novel concepts for potential integration into the KB, and *iii)* to ascertain the role of limited manual intervention in evaluating the completeness and coverage of the KB. An overview of the architecture is shown in Figure 1, while each module is presented in the following sections.

3.1 Knowledge Base-to-Prompt (*KPrompt*)

The first module involves the development of a Knowledge Base-to-Prompt strategy (*KPrompt*), which serves as the bridge between the lexical knowledge stored in the KB and the queries posed to LLMs. This strategy aims to convert the structured information within the KB into prompts that effectively capture the nuances and intricacies of the underlying semantic content. The objective is to enable LLMs to generate responses that align with the pre-existing knowledge stored in the KB, thus addressing the fundamental question of whether LLMs can proficiently deliver concepts consistent with the KB. For example, if a resource holds the information that x is connected with y through a semantic relation r (or, more formally, $r(a, b)$), then a generic template prompt for extracting b-candidates could be the following:

`Given the relation` r `with the specific meaning` $< r\_description >$`, which concepts (like` b`) might be also connected to` a `through` r`?`

Depending on the kind of knowledge encoded in the target KB, this template may be adapted in different ways and through prompting strategies such as zero- and one-shot, which are defined later on in the experimental sections.

3.2 Automatic Evaluation (*AutEval*)

The second module focuses on an intrinsic evaluation of the LLM-based extension via knowledge masking (*AutEval*), assessing the capacity of LLMs to obtain both correct and novel knowledge by first masking existing semantic units in the source knowledge and then asking LLMs to generate possible candidates (see Section 3.1 example). By systematically matching the generated candidates with the original KBs, we assess the LLM's capability to generate correct and existing information, as in the following example:

`Here is a list of candidates to connect with` $< a >$ `through` $< r >$`:` x`,` y`,` z

By checking the presence of x, y and z in the source KB (i.e. specifically of $r(a, x)$, $r(a, y)$ and $r(a, z)$), it will be possible to give some answers related to the first research question *RQ1*.

3.3 Human-in-the-loop Strategy (*HLoop*)

The third module incorporates a human-in-the-loop strategy (*HLoop*) to evaluate the novel LLM-extracted knowledge not covered by the source KB, aiming to answer the nuanced question of whether LLMs can effectively extend the KB and, simultaneously, serve as a means to verify its completeness. In particular, human evaluators, through limited manual intervention, are asked to assess the relevance and accuracy of the novel LLMs-generated knowledge.

Continuing with the example of Section 3.2, if $r(a, y)$ and $r(a, z)$ are found not to be included in the source KB, a focused manual examination of such new content may be conducted to evaluate their accuracy.

4 Experimentation

In this section, we detail the experiment settings, i.e. the implementation of the three modules (*KPrompt*, *AutEval* and *HLoop*) on three knowledge bases: Semagram (Leone et al., 2020), MultiAligNet (Grasso et al., 2022), and ConcepNet (Speer et al., 2017). The selection of these KBs has been carefully done by considering features such as scale and complexity of the encoded knowledge. By experimenting on this diversity, we aim to highlight insights and challenges under a reliable lens.

All the code for our experiments is openly available at https://anonymous.4open.science/r/LLM-Semagram-2C44/.

4.1 Prompting Strategy

KBs encapsulate complex real-world information by codifying semantic relationships, presenting a challenge for LLMs, which are typically tailored to process natural language. No standard prompting method yet exists for repurposing KB data to align with LLMs' textual processing. Our *KPrompt* methodology transfigures KB data into structured prompts for LLMs, however we do not propose *KPrompt* as a definitive standard but rather as an innovative step towards bridging the gap between KB-based data and language model processing.

Within the plethora of prompt engineering methodologies present in literature, we selected those that do not require an interactive dialogue with a LLM and are considered state of the art: Zero-Shot prompt (Kojima et al., 2022), and Few Shot prompt (Min et al., 2022; Touvron et al., 2023). In our prompts, we instruct LLMs to return 10 concepts in order to align them with the automatic evaluation in Section 5.1.

The output generated by a LLM typically consists of plain text that enumerates various concepts. To isolate these concepts (or entities), we employ regular expressions, which serve as a necessary step due to the model's potential to "hallucinate" - that is to append extraneous descriptions to the actual list of concepts. To address this, we crafted the following regular expression "$\backslash b \backslash w + \backslash b$". We also experimented with a simpler one, $\backslash w+$, but it yielded sub-optimal results across different KBs.

4.2 Knowledge Bases

In this section, we overview the KBs chosen for experimentation.

4.2.1 Semagram

The Semagram KB, introduced by Leone et al. (2020), boasts a versatile structure that captures the semantics of a given concept through a slot-filler representation. The current version encompasses over 300 concepts and 26 slots (i.e., semantic relationships). Each concept is also interconnected with other resources, e.g. BabelNet (Navigli and Ponzetto, 2010). Following (Ventrice and Siragusa, 2023), we observed that these descriptions adhered to straightforward ontology relations; for example, the *material* slot could be translated as "*can be made of*". Consequently, we opted to craft simple sentences, each posing a criterion to the LLM. Each criterion was then coupled with all its associated fillers. Subsequently, we devised a concise prompt:

```
Provide a list of 10 words that satisfy the condition.
Desired output: comma-separated list of words
Condition: can be made of wood
```

Here, *condition* encompasses the textual interpretation of the corresponding slot. This prompt structure serves as a streamlined and effective means to elicit targeted responses from the LLM based on the semantics encoded in Semagram.

4.2.2 MultiAligNet

The MultiAligNet KB, introduced by Grasso et al. (2022), constitutes a recently-developed lexical-semantic resource constructed using plain textual information gathered from several corpora in multiple languages. It encompasses knowledge across 1,047 noun concepts called *heads* and it results in 21,514 interconnected concepts. It is also linked to WordNet (Fellbaum, 2020) and BabelNet (Navigli and Ponzetto, 2010) synsets. In a simplified depiction, its internal framework resembles a KG comprising three primary node types — *noun*, *verb* and *adjective* nodes, alongside two distinct relationship types — paradigmatic and syntagmatic. Our experiment centered on the latter category, formulating prompts such as the following:

```
Provide a list of 10 English nouns related to the concept “shape, form, configuration” in the form of a comma-separated list of lowercase lemmas.
Examples: solubility, mean, packing, weight, load, color, size, style, art
```

4.2.3 ConceptNet

ConceptNet, introduced by Speer et al. (2017), serves as a multilingual KB that captures the connections and common-sense relationships among words. The inclusion of words and relationships stems from diverse sources, ranging from crowd-sourced inputs to expert-generated content. The dataset boasts more than 21M edges and over 8M nodes, with the English vocabulary alone comprising around 1,5M nodes. ConceptNet is characterized by two fundamental types of relations: symmetric relations and asymmetric relations. In particular, we focused on *UsedFor* (symmetric) and *RelatedTo* (asymmetric).

We then designed a straightforward prompt that receives a concept as input and instructs the LLM to identify 10 concepts that possess either a "*related to*" or "*used for*" relation with the given concept. An example of prompt ("*used for*") is as follow:

```
Given the concept ’car’, list 10 concepts for which ’car’ is used for, in the form of a comma-separated list.
```

4.3 LLMs Selection

Among all the different types of openly available LLMs, one way to select the optimal model is through Open LLM Leaderboard, a widely recognised LLM competition list. At the time of the selection of the model, the LLama-2 architecture

(Touvron et al., 2023) and the models fine-tuned from its associated weights were the highest ranked. The models are filtered by: pre-trained, fine-tuned on domain-specific datasets and chat models. The second and third categories are both derived from a fine-tuning on the first one. Another sub-distinction that we argue being ever so important in the current LLM panorama are storytelling fine-tuned models (Xie et al., 2023). One goal of this paper is also to discover if these models can enhance the capabilities to carry out the task under study.

For our purpose, we used three principles for LLMs selection: *i)* State-of-the-art for their respective categories at the time of selection. These were selected via an average score over different benchmarks for language capabilities of LLMs: ARC (Chollet, 2019), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2020), TruthfulQA (Lin et al., 2022), and WinoGrande (Sakaguchi et al., 2021). *ii)* With more than 30B parameters. This is justified by empirical evidence suggesting that larger language models tend to outperform smaller ones across various language tasks. *iii)* Pertaining to the three categories illustrated in Section 4.3.

For the aforementioned reasons, our choices fell on: *i)* **Yi-34B**: a model trained from scratch with the LLama-2 architecture; *ii)* **Tulu-2-70B**: a model that combines instruction and RLHF tuned chat models on a mix of publicly available, synthetic and human-created datasets; *iii)* **Aetheria-L2-70B**: a model specifically tailored for storytelling, that combines Euryale v1.3 base with the DPO training of the Tulu v2 model, and the GOAT Storytelling model. The LimaRP v3 QLora was then added (RoyalLab, 2023).

Each model was tasked to process the full set of prompts for each KB, with the sampling parameters configured to a top-p (nucleus) sampling value of 0.95, temperature of 0.4, PagedAttention enabled (Kwon et al., 2023), and a maximum token limit of 100. We used AWQ 4-bit quantization (Lin et al., 2023) to reduce memory utilization without losing language capability (Yao et al., 2023).

5 Evaluation

Within our framework, we assess two key aspects: *i)* the proficiency of the LLMs in accurately extracting verified knowledge from the KB (*AutEval*), and *ii)* the extent to which novel knowledge is extracted that was not originally encoded in the KB (*HLoop*), through manual annotation.

5.1 The *AutEval* process

We adopted standard evaluation metrics to assess the performance of LLMs. Formally, let p represent a prompt from the set P, $C_p = \{c_1, c_2, \ldots, c_n\}$ denote the list of concepts returned by the LLM, $C_k = \{c_1, \ldots, c_k\}$ the set of the first k concepts of C_p, and $K_p = \{k_1, k_2, \ldots, k_m\}$ denote the list of concepts existing in the KB and related to the prompt. The metrics are defined as follows: **Precision@K:** proportion of the returned items in the top-k (C_k) that are actually relevant; **Recall@K:** proportion of relevant items found in the top-k recommendations (K_q); and **F-Measure@K:** the harmonic mean between Precision and Recall. We also provide an asymptotic Recall value based on truncating the concept list to 10 items in the prompt.

5.1.1 Semagram

Table 1 presents the performance scores for both zero-shot and one-shot prompts across the three LLMs. The scores in the table reveal that Tulu outperforms the other models. Additionally, the performance disparity between zero-shot and one-shot prompts is marginal (0.41 percentage points for F1@10).

Prompt		P@10	R@10	F1@10
Zero-shot	Aetheria	8.73	44.00	14.57
	Tulu	**9.59**	**46.36**	**15.89**
	Yi	5.82	25.30	9.46
One-shot	Aetheria	7.52	36.21	12.45
	Tulu	**9.42**	**43.36**	**15.48**
	Yi	5.9	28.11	9.75

Table 1: Results on Semagram for Precision, Recall and F1-Measure at 10.

Figure 2 illustrates the Precision and Recall scores for varying values of k: 1, 2, 5, and 10. A notable trend is that at lower values of k, one-shot prompts outperform the zero-shot in both metrics. This trend reverses when $k > 5$.

5.1.2 MultiAligNet

We constructed distinct prompts for each syntactic category—*noun*, *verb*, and *adjective*—to evaluate potential performance discrepancies across these types. Their outcomes are detailed in Table 2. Our findings suggest that model efficacy varies significantly with syntactic category; specifically, Aetheria and Tulu demonstrate superior Precision and Recall for *noun* nodes, outstripping *verb* and

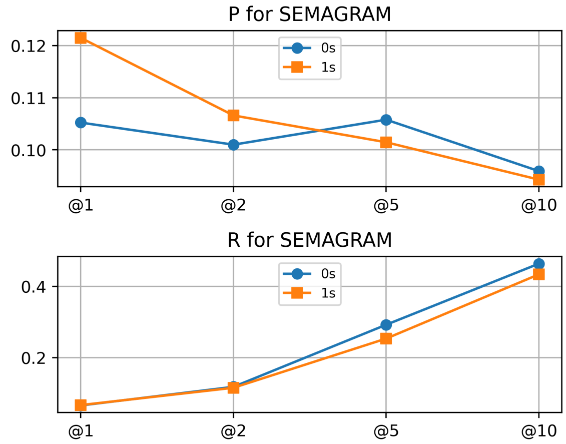

Figure 2: Precision and Recall on Semagram as k increases (best performing model). The asymptotic Recall value (at 10) is 0.98.

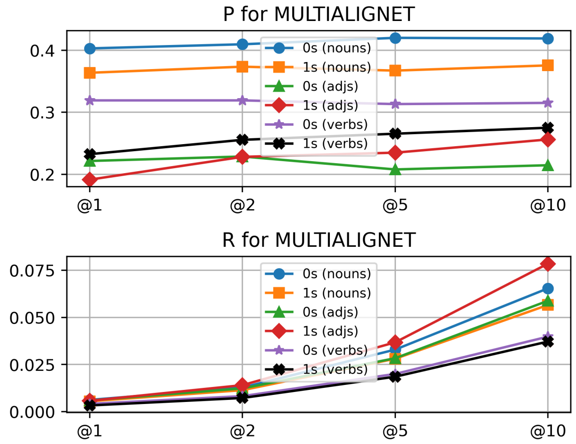

Figure 3: Precision and Recall on MultiAligNet as k increases (best performing models). The asymptotic Recall value (at 10) for $nouns$ is 0.17, for $adjectives$ is 0.32, for $verbs$ is 0.14.

adjective nodes. Conversely, for *verb* and *adjective* nodes, Aetheria and Yi lead in zero-shot and one-shot settings. Overall, the zero-shot prompting strategy on Aetheria demonstrates the most efficacy, except for *adjective* nodes.

Prompt		P@10	R@10	F1@10
Zero-shot (Nouns)	Aetheria	**41.89**	**6.54**	**11.30**
	Tulu	39.47	6.00	10.40
	Yi	12.21	1.78	3.11
Zero-shot (Adjs)	Aetheria	**21.45**	**5.88**	**9.23**
	Tulu	16.16	4.36	6.87
	Yi	4.74	1.24	1.97
Zero-shot (Verbs)	Aetheria	**31.5**	**3.98**	**7.07**
	Tulu	30.20	3.50	6.27
	Yi	8.58	1.02	1.83
One-shot (Nouns)	Aetheria	36.83	5.52	9.59
	Tulu	**37.55**	**5.66**	**9.84**
	Yi	26.23	4.25	7.33
One-shot (Adjs)	Aetheria	20.54	4.94	7.97
	Tulu	22.32	5.77	9.18
	Yi	**25.61**	**7.84**	**12.00**
One-shot (Verbs)	Aetheria	27.29	3.20	5.73
	Tulu	26.86	3.29	5.86
	Yi	**27.5**	**3.72**	**6.55**

Table 2: Results on MultiAligNet for Precision, Recall and F-Measure at 10.

Figure 3 presents the performance metrics across varying values of k. The observed trends align with those reported in Section 5.1.1, albeit with notable distinctions. For *verb* and *adjective* nodes, there is a minor but consistent enhancement in Precision.

5.1.3 ConceptNet

As discussed in Section 4.2.3, we focused on *RelatedTo* and *UsedFor* relationships. Table 3 presents the results, where Tulu, employing a one-shot prompt strategy, achieves the highest scores. The performance gap between zero-shot and one-shot is about 1.03 percentage points. Differently, Yi demonstrated a notably poorer performance. Finally, the one-shot strategy aided the models in comprehending the task, thereby enhancing their ability to retrieve more accurate concepts.

In Figure 4, Tulu consistently shows the highest Precision and Recall values for the *RelatedTo* relationship across different values of k. The Recall scores remain comparable up to $k = 2$, with a slight improvement for the one-shot prompt beyond this point. Differently, the *UsedFor* relationship exhibits lower performance overall, with a peak on Precision at $k = 2$.

5.2 The *HLoop* process

Our comprehensive automatic evaluation indicated a consistent pattern of moderate to low Precision and Recall scores across both fine-tuned (Aetheria, Tulu) and non-fine-tuned (Yi) models. Such results pointed out the necessity for an additional layer of scrutiny. Therefore, we structured a manual evaluation to understand whether the unsatisfactory scores stemmed from model errors or the generation of novel data absent from the KBs.

To this end, we selected a sample of 300 prompt outputs, from the best performing models, referring respectively to each of the three employed

Prompt		P@10	R@10	F1@10
Zero-shot (RelTo)	Aetheria	19.62	10.68	13.83
	Tulu	**20.25**	**10.38**	**13.72**
	Yi	9.37	4.60	6.17
Zero-shot (UsedFor)	Aetheria	6.79	3.29	4.43
	Tulu	**8.7**	**3.98**	**5.46**
	Yi	1.92	0.8	1.13
One-shot (RelTo)	Aetheria	20.8	11.43	14.75
	Tulu	**22.18**	**11.96**	**15.54**
	Yi	8.16	4.07	5.43
One-shot (UsedFor)	Aetheria	7.06	3.43	4.62
	Tulu	**8.46**	**4.09**	**5.51**
	Yi	4.20	1.80	2.50

Table 3: Results on ConceptNet for Precision, Recall and F-Measure at 10.

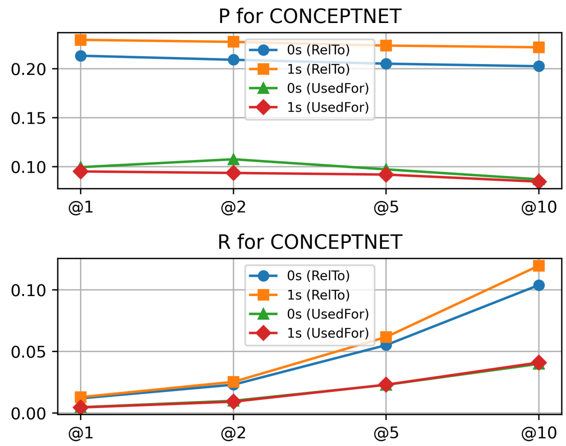

Figure 4: Precision and Recall on ConceptNet as k increases (best performing model). The asymptotic Recall value (at 10) for $RelatedTo$ is 0.67, for $UsedFor$ is 0.56.

KBs. We then asked three annotators to examine the 900 prompt outputs and to determine if a generated concept is related to the target one within the predefined relationship. The potential verdicts for each entry were categorized as correct, incorrect, or misspelled, with the latter specifically denoting any grammatical inaccuracies introduced by the models.

We used Fleiss (1971)'s kappa (F-κ) and Randolph (2005)'s multirated kappa (R-κ) to evaluate the Inter Annotator Agreement (IAA). Both metrics provide a lower and upper bound on the IAA. For Semagram, we obtained an F-κ of 0.43 – moderate agreement – and an R-κ of 0.60 – substantial agreement. MultiAligNet has an F-κ of 0.51 – moderate agreement – and an R-κ of 0.64 – substantial agreement. Finally, ConceptNet obtained the lowest agreement scores, having an F-κ of 0.33 – fair agreement – and an R-κ of 0.5 – moderate agreement.

Figure 5 shows the results of the annotation. Both Semagram and MultiAligNet have almost the same amount of correct and incorrect concepts, with a small difference on the misspelled (8 vs 6). Having such a small pool of misspelled concepts demonstrates that our prompting methodology does not elicit word hallucinations. ConceptNet, instead, has a large amount of incorrect concepts (179 incorrect vs 115 correct). These results, discussed in Section 5.3, show that LLMs "*prefer*" more fine-grained and specific relationships, whereas they hallucinate on more generic and abstract relationships.

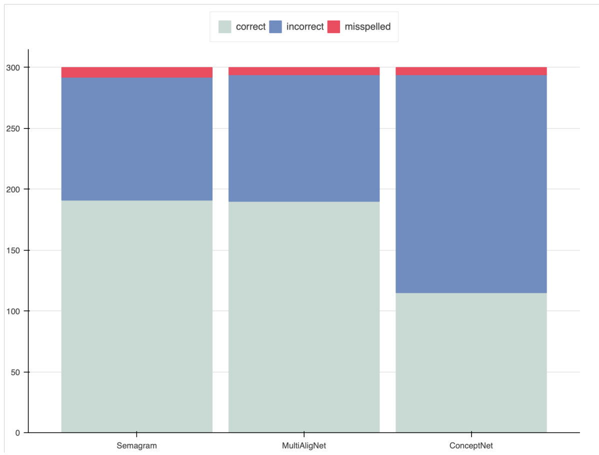

Figure 5: The ratio of correct, incorrect, and misspelled concepts on the three KBs.

5.3 Discussion

Throughout the evaluation phases, we noticed further interesting patterns that serve as additional contributions of the paper.

5.3.1 KB Size and Recall Relationship

From the observations gathered across varying sizes of KBs, we noticed an increase of Recall rates over the KB sizes. This phenomenon can be attributed to bigger data coverage within the KB (i.e., the greater the volume of entries within a KB, the higher the probability that it encompasses information pertinent to a broader spectrum of queries (Kandpal et al., 2023)).

5.3.2 Fine-tuned Models are Better

Another confirmed hypothesis regards the relationship between LLMs fine-tuning and performance in generating new and correct data for the KBs. Both Tulu and Aetheria performed 2 and 2.1 times better than the standard pre-trained Yi, given their Recall scores on the three KBs. Finally, the storytelling model, Aetheria, performed slightly better.

5.3.3 LLMs Cannot Improvise

LLMs could only recall up to 40% of the information within the tested resources, often falling short of this benchmark, verging on null accuracy. This underscores a noteworthy constraint in their capabilities. Even with guidance (i.e. using a one-shot prompt), their low recall capability does pose three possible hypothesis: *i)* LLMs cannot infer semantic relationships; *ii)* the given prompts are not accurate enough to give an explanation of the task to the model, and so contextual information may be missing. LLMs can struggle with understanding and maintaining context, especially if the KB is complex and the semantic relationships are too general; *iii)* LLMs might misunderstand the meaning of terms within the KB and thus fail to recall relevant information that depends on a different interpretation of these terms.

All of the above can be in fact true knowing that *i)* LLMs do not "understand" semantics because they have no formal grounding or theory of mind (Pavlick, 2023; Ullman, 2023); *ii)* LLMs are heavily context dependent (Shi et al., 2023); *iii)* when unguided, they fail to resolve ambiguity in language most of the time (Zhao et al., 2021; Liu et al., 2021; Zhang et al., 2022).

5.3.4 Results from Manual Analysis

Diving deeper into the results, it is peculiar to see some divergency. On one hand, on KBs that can be considered "fine-grained" (e.g. Semagram, MultiAligNet), LLMs seem to perform better and generate new and usable knowledge. On the other, common-sense KBs seem to heavily challenge LLMs because of the aforementioned missing context and semantic ambiguity.

This brought to a vast amount of confabulated content, mostly dependent to directionality problems in semantic relationships. This is due to the "Reversal Curse", discovered in (Berglund et al., 2023), that states "*if a model has been trained on a sentence of the form "A is B", it will not automatically generalize to the reverse direction "B is A" whatsoever*".

6 Conclusions

This study embarked on an exploration of Large Language Models to reconfigure lexical-semantic information, leveraging three existing resources. Our objectives encompassed assessing their efficacy and precision in this endeavor, alongside examining their ability to autonomously enrich these resources, as verified through human evaluation. Implicitly, our methodology also scrutinized the comprehensiveness of the original resources.

In synthesis, LLMs, in their current state of training, evince notable limitations in the domain of lexical semantics, extending beyond the realm of prompt variability and output alignment challenges (Kim et al., 2023). In resources teeming with rich contextual nuances (thus not solely reliant on decontextualized x-rel-y relations), LLMs manifest a pronounced capacity for generating novel knowledge. Finally, the moderate concordance among human evaluators in assessing LLM outputs underscores significant inadequacies within the resources themselves. The encoded information, or its attempted encoding, appears markedly unstable, subjective, and frequently incomplete, thereby signifying a pressing imperative for further refinement and augmentation.

6.1 Future Work

Although our paper exposes the limits of LLMs in generating data for existing KBs, recent work suggests that, while these models may not be suitable for truth-telling, they excel at revising incorrect data and identifying mistakes (Gou et al., 2023; Tyen et al., 2023). One potential future research direction is to modify our prompt engineering methodology from a purely zero/one-shot approach to a pipeline that incorporates both prompting and revision. Another research direction could be the use of RAG (Retrieval Augmented Generation) and KEG (Knowledge Enriched Generation) (Lewis et al., 2020; Yu et al., 2021; Gao et al., 2023) to enhance the context capabilities of these models. Retrieving current KB data instead of explaining through examples might be the key to unlock their capabilities. Finally, another direction involves the creation of a dataset of probing questions to assess the ability of LLMs in generating accurate and coherent data for KBs, serving as a benchmark to compare LLM performance and advance the development of more sophisticated models for this task.

Limitations

Our study introduces a novel approach to enhancing lexical semantic KBs using LLMs, yet it comes with several limitations that warrant attention. The methodology is deeply entwined with the capabilities of LLMs, meaning that any intrinsic limitations, such as biases in training data or a lack of deep context understanding, are directly reflected in the quality of our generated knowledge. The complexity of semantic relationships within the KB also significantly influences LLM performance, with fine-grained KBs yielding better results compared to those with more abstract, common-sense relationships.

Furthermore, our focus on English-language resources limits the applicability of our findings to KBs in other languages, particularly those with complex morphology or unique syntax. The validation process also revealed a moderate level of agreement among human annotators, highlighting the subjective nature of interpreting LLM outputs, which could introduce inconsistencies in the assessment of knowledge completeness and validity.

Additionally, while human evaluation is critical for ensuring the quality of LLM outputs, it is not scalable and requires substantial manual effort, posing a challenge for larger KBs or ongoing maintenance. The issue of LLMs potentially generating plausible but incorrect information, known as "hallucinations" or using a better word, "confabulations"[1], persists despite our efforts to minimize it through strategic prompting. Finally, our study's success hinged on the meticulous crafting of prompts, a process lacking standardized best practices, and remains a significant challenge in eliciting consistently accurate and relevant responses from LLMs.

Ethics Statement

The experiment was designed with the idea of providing beneficial knowledge and not harm any individual or group. Our primary goal was to develop a methodology for expanding the coverage of existing lexical KBs using Large Language Models (LLMs). We recognize that the use of LLMs carries potential risks and ethical considerations, and we have taken steps to mitigate these risks throughout our research. For example, using open weights LLMs and open sourcing our software helps the community understand the concept of risk mitigation and reproducibility in experiments.

We recognize that the use of LLMs can have environmental and social impacts. We have made efforts to minimize the environmental impact of our research by optimizing our code and using energy-efficient hardware. By using AWQ quantization, we allowed the models to run on one A100 GPU. The estimated working hours of the single GPU was of 20 hours, for a CO_2 emission of 2.8 kg. In our commitment to offset these emissions, we have initiated the establishment of a forest through Treedom[2]. As an initial endeavor, we have planted a tree, uniquely identified with the code YMZ-6K66.

[1] https://www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate/

[2] https://www.treedom.net/

References

Gabor Angeli, Michael Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *NeurIPS Workshop on Attributing Model Behavior at Scale*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Muhao Chen, Yaxin Tian, Mohan Yang, Puneet Mathur, and Yaxin Chen. 2021. Knowledge graph embedding: A survey of approaches and applications. *ACM Computing Surveys (CSUR)*, 54(5):1–35.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1856–1869, Dubrovnik, Croatia. Association for Computational Linguistics.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

Christiane Fellbaum. 2020. *WordNet*. Princeton University Press.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zhibin Gou, Zhihong Shao, Yeyun Gon g, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing.

Francesca Grasso, Vladimiro Lovera Rulfi, and Luigi Di Caro. 2022. Multialignet: Cross-lingual knowledge bridges between words and senses. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 36–50. Springer.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

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.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo, and Minjoon Seo. 2023. Aligning large language models through synthetic feedback. In *EMNLP*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Valentina Leone, Giovanni Siragusa, Luigi Di Caro, and Roberto Navigli. 2020. Building semantic grams of human knowledge. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2991–3000, Marseille, France. European Language Resources Association.

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

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*.

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

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Dabin Min, Kaeun Kim, Jong Hyuk Lee, Yisak Kim, and Chang Min Park. 2022. RRED : A radiology report error detector based on deep learning framework. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 41–52, Seattle, WA. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041.

Erika Petersen and Christopher Potts. 2023. Lexical semantics with large language models: A case study of English "break". In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 490–511, Dubrovnik, Croatia. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Mike Lewis, and Sebastian Riedel. 2019. Language models as knowledge bases? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Simon Razniewski and Gerhard Weikum. 2021. Zero-shot open knowledge base question answering. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pages 1095–1104.

RoyalLab. 2023. royallab/Aetheria-L2-70B · Hugging Face — huggingface.co. `https://huggingface.co/royallab/Aetheria-L2-70B`. [Accessed 01-02-2024].

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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

Gladys Tyen, Hassan Mansoor, Peter Chen, Tony Mak, and Victor Cărbune. 2023. Llms cannot find reasoning errors, but can correct them! *arXiv preprint arXiv:2311.08516*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Laura Ventrice and Giovanni Siragusa. 2023. Enhancing semantic resources via large language models. *GENERAL'23: GENerative, Explainable and Reasonable Artificial Learning Workshop 2023*.

Duyu Vo and Ebrahim Bagheri. 2017. Extracting structured data from natural language with semi-structured knowledge. *arXiv preprint arXiv:1710.10723*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *International Conference on Natural Language Generation*.

Zhewei Yao, Cheng Li, Xiaoxia Wu, Stephen Youn, and Yuxiong He. 2023. A comprehensive study on post-training quantization for large language models. *arXiv preprint arXiv:2303.08302*.

Wenhao Yu, Meng Jiang, Zhiting Hu, Qingyun Wang, Heng Ji, and Nazneen Rajani. 2021. Knowledge-enriched natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 11–16.

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

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.