Where and How as Key Factors for Knowledge-Enhanced Constrained Commonsense Generation

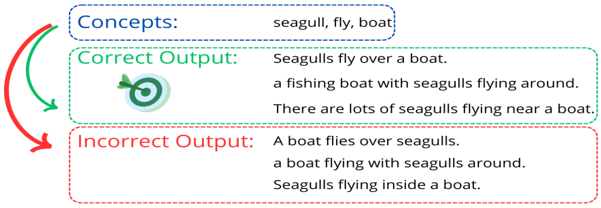
Iván Martínez-Murillo, Paloma Moreda, Elena Lloret
Department of Language and Computing Systems, University of Alicante,
Alicante, Spain, 03690.

Abstract

This paper addresses a key limitation in Natural Language Generation (NLG) systems: their struggle with commonsense reasoning, which is essential for generating contextually appropriate and plausible text. The study proposes an approach to enhance the commonsense reasoning abilities of NLG systems by integrating external knowledge framed in a constrained commonsense generation task. The paper investigates strategies for extracting and injecting external knowledge into pre-trained models, specifically BART and T5, in both base and large configurations. Experimental results show that incorporating external knowledge, extracted using a simple strategy, leads to significant performance improvements, with the models achieving 88% accuracy in generating plausible and correct sentences. When refined methods for knowledge extraction are applied, the accuracy further increases to 92%. These findings underscore the crucial role of high-quality external knowledge in enhancing the commonsense reasoning capabilities of NLG systems, suggesting that such integration is vital for advancing their performance in real-world applications.

1 Introduction

The rapid advancement of Natural Language Generation (NLG) systems has significantly transformed the field of Artificial Intelligence (AI). Modern NLG systems are capable of producing text in a very similar way humans would do, with high coherence and fluency (Sepúlveda-Torres et al., 2025). Moreover, these systems demonstrate impressive performance across a wide range of tasks (Liang et al., 2024). Despite these achievements, one critical area where such models continue to fall short is commonsense reasoning for generating plausible and fluent sentences (Miró Maestre et al., 2025; Seo et al., 2024).



Seagulls fly over a boat.
a fishing boat with seagulls flying around.
There are lots of seagulls flying near a boat.

A boat flies over seagulls.
a boat flying with seagulls around.
Seagulls flying inside a boat.

Figure 1: Constrained commonsense text generation task.

Commonsense knowledge plays a fundamental role in human communication, enabling individuals to make inferences without the need for explicit elaboration. This intuitive ability is crucial for effective and contextually appropriate communication. In the context of NLG, the absence of commonsense reasoning can result in outputs that, although plausible, often lack fluency or contextual appropriateness, as illustrated by the examples in the red box in Figure 1.

A common way to address this limitation is by integrating external knowledge into NLG models (Jiang et al., 2024). In this context, external knowledge refers to information from knowledge bases that helps enhance the model's understanding. By using such knowledge, NLG systems can produce sentences that are more plausible and better aligned with the given context.

This paper investigates strategies to improve commonsense reasoning in NLG by incorporating external knowledge into the constrained commonsense NLG task. The goal is to generate a plausible sentence that includes a given set of input words as input. Figure 1 illustrates an example of this task, highlighting sentences in the red box that are syntactically correct but not semantically valid. This task is challenging because NLG models often produce fluent text, but without a deep understanding of what they generate. Adding external knowledge can help models reason more effectively, resulting in more coherent and contextually accurate outputs. This study focuses on addressing two main research

Proceedings of Recent Advances in Natural Language Processing,pages 694–703
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-080

questions:

- *Where should external knowledge be injected into pre-trained models?* To answer this, we analyze the most appropriate stage for knowledge integration—whether during fine-tuning, inference, or both.

- *How can external knowledge be effectively retrieved?* We conduct a study on different extraction techniques and assess their impact on the quality of the generated text.

2 Related Work

Constrained Commonsense Text Generation: Constrained Text Generation is a subtask of NLG that focuses on producing text under specific restrictions. These constraints can be lexical, length-based, semantic, syntactic, or stylistic (Zhou et al., 2023). Among the lexical-constrained text generation tasks, the CommonGen (Lin et al., 2020) task emerges to test the commonsense reasoning ability of generative NLG models. Given three keywords as input, the system must generate a plausible sentence containing those words.

The best approaches addressing this task have relied on the one hand training a retriever of documents to identify relevant candidate sentences that are then used for the generation process as a context (He et al., 2022; Yu et al., 2022b), and on the other hand, filtering the knowledge to input the model and applying machine learning methods for training those models (Li et al., 2021).

Knowledge Enhanced Approaches: Knowledge-enhanced NLG aims to improve the relevance and accuracy of generated text by incorporating external knowledge (Yu et al., 2022a). This knowledge can come from three main sources: structured knowledge bases, knowledge graphs, and grounded text. Knowledge bases and graphs are often organized as subject-predicate-object triplets. In graphs, these elements are further connected in a tree-like structure, which improves their interrelation. Grounded text refers to textual information collected from online sources.

Different works proposed methods for enhancing NLG systems with external knowledge. KG-BART is a NLG system which is built on the BART model by replacing its standard Transformer with a Knowledge Graph-Augmented Transformer. This change allows the model to use grounded knowledge graphs more effectively through a graph attention mechanism. Similarly, Wang et al. (2021) enhances pre-training by retrieving related sentences from external corpora and uses a trainable retriever during fine-tuning to improve results.

Despite promising outcomes, knowledge-enhanced NLG still faces challenges (Hu et al., 2023). These systems often depend on static knowledge, which can become outdated. Additionally, if retrieved information is not properly aligned with the context, it may introduce errors or hallucinations. Finally, evaluating the true benefit of external knowledge also remains difficult due to the lack of faithful evaluation metrics.

3 Knowledge-enhanced NLG Strategies

In this section, we propose a three-stage approach for addressing the task of lexically constrained commonsense NLG. The proposed approach is illustrated in Figure 2.

1. **Knowledge extraction**: We dynamically retrieve knowledge from an external knowledge base by matching each word candidate to be included in the generated sentence with relevant entries in the knowledge base. We experiment with different methods to retrieve the knowledge.

2. **Prompt Engineering**: We combine the extracted knowledge with the corresponding dataset entry using a prompt engineering method based on knowledge contextualization. In this approach, each keyword that needs to be included in the generated sentence is matched with the related extracted knowledge. The obtained relations are then added to the prompt to provide context. The prompt starts with an initial premise *"Generate a short sentence using the following words"*, and this is followed by the concepts to be included in the generated sentences and the retrieved most-common relation for those words.

3. **Knowledge Injection**: A pre-trained model is then fine-tuned on the dataset enriched with this external knowledge. We conduct a series of experiments to determine the optimal stage for knowledge injection (during fine-tuning, during the models' inference, or in both). As a consideration, incorporating knowledge during the inference step ensures that the information remains as up-to-date as possible.

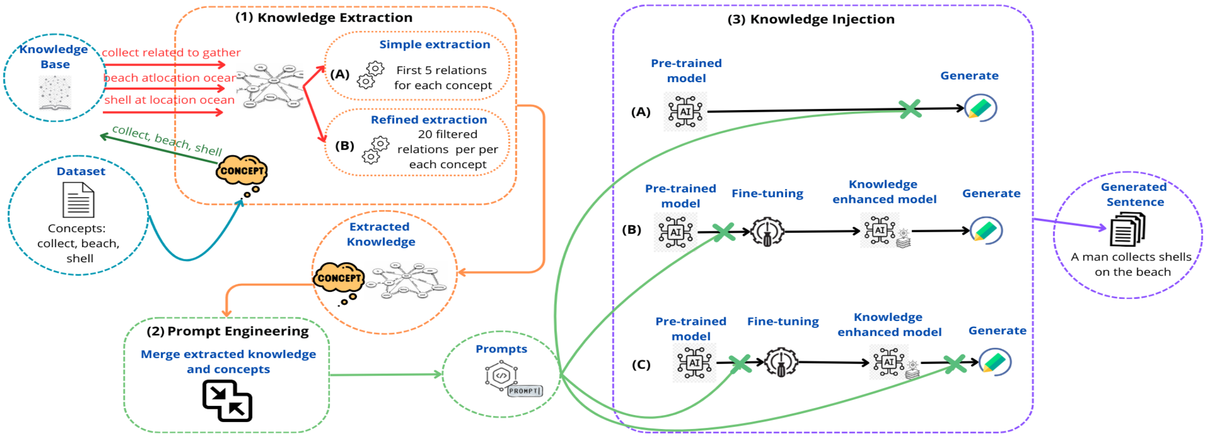


Figure 2: Proposed approach to enhance commonsense in NLG systems

4 Resources

This section describes the experimental setup for evaluating the effectiveness of augmenting pre-trained models with external knowledge.

4.1 Dataset

We frame our generation task within the CommonGen shared task (Lin et al., 2020), using its corpus for our experiments. The task aims to generate coherent sentences that describe everyday scenarios with a given set of keywords, testing the ability of NLG systems to produce commonsense outputs while including those keywords.

The dataset is split into training, validation, and testing subsets, with each keyword set associated with up to three reference sentences as targets. Dataset details are summarized in Table 1.

Corpus	*Train*	*Dev*	*Test*
Concept sets	32 651	993	1 497
Target sentences	67 389	4 018	N/A

Table 1: CommonGen Dataset distribution. N/A indicates that no records are available.

Since the reference sentences for the CommonGen test subset are not publicly available, we used the validation subset as our test subset, to be able to perform an automatic evaluation of the generated sentences. We also excluded one sentence from the validation subset due to the word "cain", which was treated inconsistently in the original dataset as "cane". This led to the validation subset consisting of 992 sentences, which composed our test subset. Furthermore, we split the original training subset into two subsets, using 90% for training and 10% for validation. This split ensures model generalization by evaluating performance on unseen data before final testing.

4.2 Knowledge Base

By leveraging external knowledge sources, NLG models can generate more coherent and contextually appropriate sentences while reducing the likelihood of hallucinations. To achieve this, we incorporated ConceptNet (Speer et al., 2017), as the external knowledge graph. ConcepNet is a widely used knowledge graph that encodes commonsense relationships between words in the form of triplets (e.g. "Word - RelatedTo - Sentence"). ConceptNet is particularly valuable for the CommonGen task, as it provides structured, human-curated associations that reflect real-world interactions. Since CommonGen focuses on generating plausible sentences describing everyday scenarios, ConceptNet enhances the fluency and realism of generated sentences.

4.3 Pre-trained Models

Pre-trained models are transformers trained on large datasets, which acquire general knowledge that can be fine-tuned for specific tasks with minimal training. Compared to LLMs, pre-trained models are smaller and more efficient, requiring fewer computational resources. Despite their smaller size, fine-tuned pre-trained models can often achieve comparable performance to LLMs in certain NLG tasks (Li et al., 2024). This efficiency makes them well-suited for scenarios with resource constraints. Based on these strengths, we have chosen to experiment with pre-trained models as the foundation of our methodology.

In the context of the CommonGen task, the approaches with the best results used BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) as foundational models[1]. As a result, we selected these models for our experiments, testing two versions of each: BART-Base and BART-Large, as well as T5-Base and T5-Large. The BART-Base model features 6 encoder-decoder layers and about 140 million parameters, while BART-Large has 12 encoder-decoder layers and approximately 400 million parameters. For T5, the Base version contains 220 million parameters, while T5-Large has 770 million parameters. Both models are encoder-decoder architectures pre-trained for text-to-text tasks.

4.4 Fine-tuning

After successfully extracting and injecting the retrieved external knowledge, the next step is to train the models to perform effectively on the CommonGen task. To achieve this, we fine-tuned the pre-trained models mentioned in Section 4.3 using the constructed inputs through the prompt-engineering step. The hyperparameter configuration used during the training process is as follows: we trained our systems for a total of 4 epochs, with a batch size of 32, a dropout rate of 0.1, a learning rate of 1e-4, and 1000 warm-up steps. All the experiments were conducted on a single NVIDIA A100 GPU.

5 Experiments & Evaluation

In this section, we outline the experiments conducted to address the research questions stated in Section 1.

5.1 Where to Inject External Knowledge?

We conducted different experiments to evaluate the impact of incorporating external knowledge in pre-trained models. As previously mentioned, we used the BART and T5 models in their base and large versions. The objective was to determine whether the inclusion of external knowledge impacts model performance and to identify where knowledge injection is most effective—during training, during inference, or both. To this end, we designed experiments under four conditions:

1. *No external knowledge*, where the models were fine-tuned on the CommonGen train dataset and the generated sentences were produced from the trained models over the development set. This experiment serves as a baseline, where no external knowledge is introduced.

2. *Knowledge in Inference*, in which we use a base model without fine-tuning it, relying only on external knowledge we extracted to assist in the generation step (path A in Stage 3 of Figure 2).

3. *Knowledge in Train*, in which our extracted external knowledge is incorporated during the fine-tuning process (path B in Stage 3 of Figure 2). The sentences are then generated using the resulting model on the CommonGen development set, without the help of the external knowledge being provided during the generation step.

4. *Knowledge in All* that combines the injection of external knowledge during both fine-tuning and inference steps (refer to Path C in Stage 3 of Figure 2).

To evaluate the experiment on determining where it is most optimal to inject knowledge, we relied on automatic metrics. Moreover, we aimed to compare our results using the same metrics employed in the CommonGen task. For this purpose, we utilized two of the evaluation metrics proposed in the CommonGen task: CIDEr (Vedantam et al., 2015) and BLEU_4 (Papineni et al., 2002). Additionally, we also evaluated the generated sentences using BLEU_1. BLEU_1 was included to specifically assess unigram precision, providing insight into basic word-level accuracy.

Moreover, we incorporate additional metrics beyond those originally employed in the CommonGen task. These supplementary metrics—ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), the Hallucination Evaluation Model (HHEM) (Bao et al., 2024), and AlignScore (Zha et al., 2023)—allow for a more nuanced assessment of various aspects of text quality. By leveraging multiple evaluation metrics, we aim to mitigate the limitations of any single metric and obtain a more holistic understanding of the model's performance.

5.2 How to Extract and Incorporate External Knowledge Effectively?

After identifying the optimal phase for injecting knowledge, we experimented with different methods of knowledge integration and examined to which extent refining the extracted knowledge

[1] https://inklab.usc.edu/CommonGen/leaderboard.html

impacts the output. We followed two different methodologies to extract the knowledge from ConceptNet:

- ***Simple Extraction***: For each keyword in the CommonGen dataset, we queried the ConceptNet API to retrieve its associated relations. In this approach, we obtained only the first five instances returned by the API without any further filtering.

- ***Refined Extraction***: We queried the ConceptNet API for each keyword in the CommonGen dataset. However, in this approach, we first analyze and select specific types of relations to ensure that the extracted knowledge is more relevant to the context of general, everyday concepts represented by the keywords. The relations that are most suitable for everyday scenarios are:
  - IsA: Indicates that A is a subtype of B.
  - RelatedTo: Denotes a general relation between A and B when the specific type of relation is undefined.
  - PartOf: Specifies that object A is a component or subset of B.
  - HasA: Suggests that B is an inherent part of A or represents a social construct of possession.
  - AtLocation: Describes A as a typical or likely location for B.
  - HasProperty: Declares that A can be characterized as having property B.

  For each of these selected relations, we extracted up to 20 instances.

To determine how refining the search for external knowledge can be beneficial and to what extent it improves the quality of the generated texts, we conducted a manual evaluation. As demonstrated in prior works (Martínez-Murillo et al., 2024; Sai et al., 2022), automatic evaluation metrics do not always align well with human judgments, particularly for free-text generation tasks where the output is not constrained to a specific style. Thus, we manually classified the generated sentences into three-level categories to obtain better insights into the evaluation: 2, 1, and 0.

- *2 - Good sentences* are grammatically and semantically correct, exhibit commonsense reasoning, and describe plausible situations.
- *1 - Regular sentences* are semantically correct and describe plausible situations, but may require improvements in grammar or clarity.
- *0 - Bad sentences* are grammatically incorrect, describe implausible situations, or fail to convey the intended semantic meaning.

6 Results and Discussion

In this section, we present the results obtained based on our two research questions of our experimentation.

6.1 Determining Where to Inject External Knowledge

To identify where the injection of knowledge is most effective, we conducted a preliminary automatic evaluation of all model versions across the different stages of knowledge injection—training, inference, or both. We employed the metrics mentioned in Section 5.1 and evaluated the generated sentences from each experiment using the reference sentences from the development set as the gold standard. The results from this preliminary evaluation are presented in Table 2.

The two pre-trained model families (T5 and BART) exhibit a similar trend. The best results in most of the automatic metrics are obtained injecting the knowledge in both training and inference stages (*Knowledge in All* experiment). This approach consistently delivers superior performance across the four model versions. Surprisingly, the experiments with *No External Knowledge* achieved a similar performance with automatic metrics than the experiments *Knowledge in All*. Conversely, lower results were obtained in the *Knowledge in Inference* experiments. The reason is that the tested models have not been trained to handle the CommonGen task, thus generating sentences that are either unrelated to the task or simply replicate the prompt we give as input. In contrast, the experiments involving fine-tuning pre-trained models on the CommonGen dataset consistently yield outputs that better align with the task requirements. Among the three experiments, *Knowledge in Train* produces the lowest performance. This can be attributed to the fact that, although the model is exposed to relevant knowledge during training, it lacks access to the appropriate contextual knowledge at inference time, leading to less accurate sentence generation.

Overall, the results demonstrate that T5 slightly outperforms BART for all models' versions. No-

Model	Model Size	Experiment	CIDEr	BLEU_1	BLEU_4	ROUGE-L	BERTScore	HHEM	AlignScore
BART	Base	No External Knowledge	0.150	0.694	**0.266**	**0.516**	**0.543**	0.567	**0.631**
		Knowledge in Inference	0.076	0.478	0.127	0.397	0.519	0.382	0.29
		Knowledge in Train	0.142	0.66	0.249	0.505	0.541	0.512	0.621
		Knowledge in All	**0.151**	**0.697**	0.265	0.515	**0.543**	**0.580**	0.622
	Large	No External Knowledge	**0.148**	0.69	**0.253**	0.506	0.543	0.594	0.602
		Knowledge in Inference	0.04	0.411	0.099	0.354	0.511	0.361	0.158
		Knowledge in Train	0.136	0.649	0.242	0.501	**0.548**	0.519	**0.636**
		Knowledge All	0.146	**0.692**	0.247	**0.508**	0.541	**0.598**	0.589
T5	Base	No External Knowledge	0.164	0.726	**0.295**	**0.543**	**0.549**	0.613	0.679
		Knowledge in Inference	0.029	0.267	0.065	0.187	0.418	0.257	0.246
		Knowledge in Train	0.155	0.695	0.288	0.534	0.545	0.602	0.619
		Knowledge in All	**0.165**	**0.728**	**0.295**	0.542	0.547	**0.623**	**0.693**
	Large	No External Knowledge	0.168	**0.734**	0.307	0.551	**0.549**	0.664	**0.677**
		Knowledge in Inference	0.06	0.436	0.141	0.376	0.477	0.533	0.317
		Knowledge in Train	0.155	0.695	0.288	0.534	0.545	0.602	0.619
		Knowledge in All	**0.169**	**0.734**	**0.314**	**0.554**	0.548	**0.671**	0.662

Table 2: Results obtained with the automatic metrics for the 992 generated sentences for each experiment. Results in bold indicate the best performance for that model size according to each metric.

tably, even the T5-base model without external knowledge achieves better performance than any of the BART experiments. A manual analysis of the generated sentences reveals that this difference can be attributed to specific instances where BART fails more often to generate plausible sentences. For example, given the keywords *"yard, kid, and play"*, BART-base without external knowledge generates the sentence, "*A kid plays a ball in a yard*", whereas T5-base produces a more syntactically and semantically refined sentence, "*A kid is playing with a ball in a yard*". However, without external knowledge, both base models sometimes generate unnatural sentences. For example, given the keywords *"body, raft, water"*, BART produces "*body in the water on a raft*", while T5 generates "*body of water on a raft*". These outputs demonstrate that the models struggle to form contextually appropriate sentences. Despite these shortcomings, T5 generally produces more elaborate and coherent sentences compared to BART. Additionally, the larger T5 model outperformed the smaller version in all experiments.

Therefore, based on the results obtained through the automatic metrics, the most effective method within this experimentation to inject the knowledge is during both the training and inference steps.

6.2 Determining How to Extract and Incorporate External Knowledge Effectively Through a Manual Evaluation

We next proceeded to analyze in greater depth how to retrieve the knowledge in our method. In this context, we expanded our experimentation by refining the concept retrieval process to identify the most pertinent relations from ConceptNet and subsequently fine-tuned the best-performing model from the experiments in Section 6.1, T5-Large, by integrating this enriched knowledge.

Moreover, to verify whether the injected knowledge enhances the quality of the task, we conducted a manual evaluation. Precisely, we manually analyzed the sentences generated by the T5-Large model without external knowledge and with knowledge injected in all phases using both knowledge extraction methods: simple and refined. Each generated sentence was ranked on a three-level scale (0, 1, or 2) mentioned in Section 5.2. The results are shown in Table 3.

Enhancing the pre-trained model with additional knowledge significantly improves its performance, as evidenced by the manual analysis of the generated sentences. The number of incorrect sentences decreases from 168 without external knowledge to 89 with the incorporation of simple knowledge, representing a reduction of 79 sentences. Furthermore, with the injection of refined knowledge, the number of incorrect sentences decreases even further to 62, amounting to a reduction of 106 sentences. This corresponds to a 47% reduction with simple knowledge and a 63% reduction with refined knowledge. Similarly, the number of sentences that are not entirely correct but also not incorrect improves significantly. This category decreases from 49 sentences without additional knowledge to 25 when the model is enhanced with simple knowledge. Moreover, with the inclusion of refined knowledge, this number decreases further to 12 sentences. This represents a 49% reduction with simple knowledge and a 76% reduction with refined knowledge. The

	T5-Large		
Score	*No Knowledge*	*Knowledge in All Simple*	*Knowledge in All Refined*
Incorrect	168	88 (-47%)	62 (-63%)
Regular	49	25 (-49%)	12 (-76%)
Correct	775	878 (+13%)	918(+19%)

Table 3: Results of the manual evaluation of the 992 generated sentences in the knowledge extraction experiment.

most substantial improvement is observed in the category of correct sentences. Without knowledge, the number of correct sentences is 775. With the inclusion of external knowledge, the model improves up to 878 correct sentences out of 992 total, achieving an 88% accuracy rate for plausible and correct outputs. This accuracy increases further with the injection of refined knowledge, resulting in 918 correct sentences—92% of the total generated sentences. Compared to the baseline model without knowledge, the approaches incorporating external knowledge achieve a 13% improvement with simple knowledge and a 19% improvement with refined knowledge.

Overall, the manual evaluation results suggest that incorporating external knowledge has a greater impact than indicated by the automatic evaluation metrics. While the differences in automatic evaluation were minimal, the manual assessment revealed a significant percentage improvement.

To conduct a detailed study of the results, we analyzed sentences that were initially incorrect but became correct after the injection of external knowledge, as well as those that were correct in the original, non-enhanced model but became incorrect in the knowledge-enhanced model.

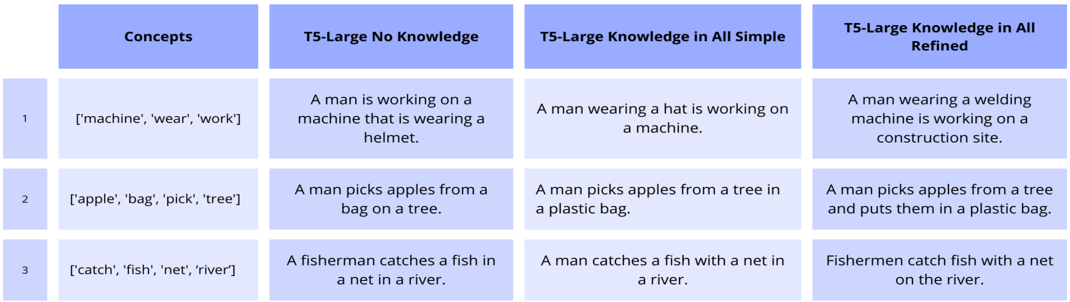
In the first scenario, 116 sentences that were generated as incorrect or partially correct without external knowledge were corrected after the injection of simple knowledge. This represents an improvement of nearly 12%. Furthermore, the improvement with the refined knowledge is bigger, being 178 the number of sentences that improved when injecting the refined knowledge.

Figure 3 presents representative examples illustrating these improvements in rows 1, 2, and 3.

- In **Row 1**, the improvement lies in the model's recognition that a machine cannot perform the action of wearing something, as this is a property of humans. Both knowledge-enhanced models correctly attribute the action of wearing to a human, each using a different object in the generated sentences.
- In **Row 2**, the implausible notion in the original sentence of placing apples from a bag onto a tree is corrected. The external knowledge helps the model understand that a bag is used for carrying items and that apples are a fruit associated with trees, leading to a more plausible scenario.
- In **Row 3**, the incorporation of external knowledge enables the model to understand that the net is the object fishermen use to catch fish, rather than the location where they catch them. Both knowledge-enhanced approaches recognize this distinction.

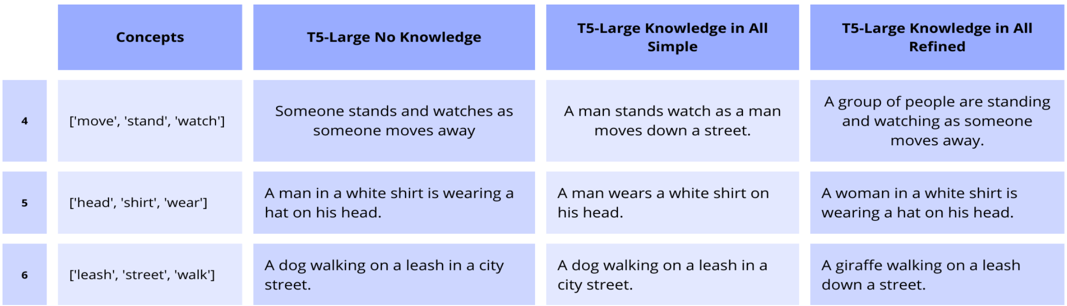
On the other hand, there are a few cases where sentences originally classified as correct without external knowledge were generated as incorrect after the injection of knowledge. This issue was observed in 12 sentences (approximately 1% of the total sentences) using the simple knowledge approach and 31 sentences (around 3%) with the refined knowledge approach. Examples of such cases can be found in rows 4, 5, and 6 of Figure 4.

- **Row 4**: In this case, the sentence generated using the simple knowledge approach is incorrect. This occurs because the external knowledge provided to the model may have introduced confusion. For the keyword "watch", the relations retrieved were solely associated with the object referring to time rather than the action of seeing. As a result, the model was unable to generate a sentence that reflected the correct meaning of the word. In contrast, the refined knowledge approach, which provided a greater number of carefully selected relations, enabled the model to interpret and use the correct sense of the verb.
- **Row 5**: In this example, the sentence generated with the simple knowledge approach is



	Concepts	T5-Large No Knowledge	T5-Large Knowledge in All Simple	T5-Large Knowledge in All Refined
1	['machine', 'wear', 'work']	A man is working on a machine that is wearing a helmet.	A man wearing a hat is working on a machine.	A man wearing a welding machine is working on a construction site.
2	['apple', 'bag', 'pick', 'tree']	A man picks apples from a bag on a tree.	A man picks apples from a tree in a plastic bag.	A man picks apples from a tree and puts them in a plastic bag.
3	['catch', 'fish', 'net', ‘river']	A fisherman catches a fish in a net in a river.	A man catches a fish with a net in a river.	Fishermen catch fish with a net on the river.

Figure 3: Samples of improved generated sentences with knowledge.



	Concepts	T5-Large No Knowledge	T5-Large Knowledge in All Simple	T5-Large Knowledge in All Refined
4	['move', 'stand', 'watch']	Someone stands and watches as someone moves away	A man stands watch as a man moves down a street.	A group of people are standing and watching as someone moves away.
5	['head', 'shirt', 'wear']	A man in a white shirt is wearing a hat on his head.	A man wears a white shirt on his head.	A woman in a white shirt is wearing a hat on his head.
6	['leash', 'street', 'walk']	A dog walking on a leash in a city street.	A dog walking on a leash in a city street.	A giraffe walking on a leash down a street.

Figure 4: Samples of worsened generated sentences with knowledge.

also incorrect. While it is plausible for a person to wear a hat or something similar on their head, it is uncommon to wear a shirt there. This confusion may have been caused by the external knowledge provided, as one of the relations for "shirt" indicated that it is worn on the upper body. This may have led the model to incorrectly associate “head” with “upper body”, resulting in an inaccurate sentence. In contrast, the refined knowledge approach enabled the model to generate a correct sentence, as it included additional relations that specified the location of a hat as the head.

- **Row 6**: In this sample, the incorrect sentence is generated by the refined knowledge approach. The error arises from the implausibility of a giraffe being on a street with a leash. This scenario is more common for dogs or pets. In this case, the external knowledge provided does not associate "leash" with a dog or pet, which fails to assist the model in generating a plausible sentence.

As seen in the examples, selecting representative and relevant relations for injecting external knowledge significantly improves sentence quality. The more precise and relevant the knowledge is, the greater the enhancement in generation it has. Conversely, incorrect or ambiguous knowledge often leads to inaccurate outputs. Thus, refining the knowledge extraction method is crucial for producing more plausible sentences.

7 Conclusions & Future Work

This study highlights the importance of incorporating external knowledge to enhance the performance of pre-trained models on NLG tasks. By integrating additional knowledge during training and inference, sentence quality and plausibility improved notably. The refined knowledge extraction method, which prioritizes relevant relations, consistently outperformed the simpler approach, resulting in a 63% reduction in incorrect sentences and increasing correct and plausible outputs to 92%.

Despite these improvements, in some cases ambiguous or incomplete external knowledge sometimes led to inaccuracies, highlighting the need for more precise knowledge extraction mechanisms.

In conclusion, the results confirm that external knowledge enhances the capabilities of pre-trained models in NLG tasks, particularly when using a refined extraction approach. Future research could explore advanced methods for dynamically selecting and injecting knowledge, as well as extending these techniques to other NLG tasks and datasets. Since this approach is language-independent—as long as knowledge bases and datasets are available in other languages—future work will also explore its multilingual potential.

Acknowledgments

The research work conducted is part of the R&D projects “CORTEX: Conscious Text Generation” (PID2021-123956OB-I00), funded by MCIN/ AEI/10.13039/501100011033/ and by “ERDF A way of making Europe”; “CLEAR.TEXT:Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities” (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and “European Union NextGenerationEU/PRTR”; “VIVES: “Pla de Tecnologies de la Llengua per al valencià” project (2022/TL22/00215334) from the Ministerio de Transformación Digital y por el Plan de Recuperación, Transformación y Resiliencia – Financiado por la Unión Europea – NextGenerationEU; and the project “NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/021)" funded by the Generalitat Valenciana. This work is also funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. Moreover, the experiments were carried out using the DGX units within the framework of project IDIFEDER/2020/003, co-funded by the Valencia Government and the European Regional Development Fund (ERDF).

References

Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

Xingwei He, Yeyun Gong, A-Long Jin, Weizhen Qi, Hang Zhang, Jian Jiao, Bartuer Zhou, Biao Cheng, Sm Yiu, and Nan Duan. 2022. Metric-guided distillation: Distilling knowledge from the metric to ranker and retriever for generative commonsense reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 839–852, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*.

Meng Jiang, Bill Yuchen Lin, Shuohang Wang, Yichong Xu, Wenhao Yu, and Chenguang Zhu. 2024. Knowledge-augmented methods for natural language understanding. In *Knowledge-augmented Methods for Natural Language Processing*, pages 23–40. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haonan Li, Yeyun Gong, Jian Jiao, Ruofei Zhang, Timothy Baldwin, and Nan Duan. 2021. KFCNet: Knowledge filtering and contrastive learning for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2918–2928, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable text generation for large language models: A survey. *CoRR*, abs/2408.12599.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Iván Martínez-Murillo, Paloma Moreda, and Elena Lloret. 2024. Analysing the problem of automatic evaluation of language generation systems. *Procesamiento del Lenguaje Natural*, 72:123–136.

María Miró Maestre, Iván Martínez-Murillo, Tania Josephine Martin, Borja Navarro Colorado, Antonio Ferrández, Armando Suárez Cueto, Elena Lloret, et al. 2025. Roadmap for natural language generation: Challenges and insights. *Procesamiento del Lenguaje Natural*, 74(0):67–79.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Jaehyung Seo, Jaewook Lee, Chanjun Park, SeongTae Hong, Seungjun Lee, and Heuiseok Lim. 2024. Ko-CommonGEN v2: A benchmark for navigating Korean commonsense reasoning challenges in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2390–2415, Bangkok, Thailand. Association for Computational Linguistics.

Robiert Sepúlveda-Torres, Iván Martínez-Murillo, Estela Saquete, Elena Lloret, and Manuel Palomar. 2025. To write or not to write as a machine? that's the question. *IEEE Transactions on Big Data*, pages 1–12.

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.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Han Wang, Yang Liu, Chenguang Zhu, Linjun Shou, Ming Gong, Yichong Xu, and Michael Zeng. 2021. Retrieval enhanced model for commonsense generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3056–3062, Online. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022a. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022b. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4364–4377, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *International Conference on Machine Learning*, pages 42602–42613. PMLR.