Zero-Shot Commonsense Validation and Reasoning with Large Language Models: An Evaluation on SemEval-2020 Task 4 Dataset

Rawand Alfugaha
College of Information Technology
Lusail University
Doha, Qatar
ralfoqha@lu.edu.qa

Mohammad AL-Smadi
Digital Learning and Online Education Office
Qatar University
Doha, Qatar
malsmadi@qu.edu.qa

Abstract

This study evaluates the performance of Large Language Models (LLMs) on SemEval-2020 Task 4 dataset, focusing on commonsense validation and explanation. Our methodology involves evaluating multiple LLMs, including LLaMA3-70B, Gemma2-9B, and Mixtral-8x7B, using zero-shot prompting techniques. The models are tested on two tasks: Task A (Commonsense Validation), where models determine whether a statement aligns with commonsense knowledge, and Task B (Commonsense Explanation), where models identify the reasoning behind implausible statements. Performance is assessed based on accuracy, and results are compared to fine-tuned transformer-based models. The results indicate that larger models outperform previous models and perform closely to human evaluation for Task A, with LLaMA3-70B achieving the highest accuracy of **98.40%** in Task A whereas, lagging behind previous models with **93.40%** in Task B. However, while models effectively identify implausible statements, they face challenges in selecting the most relevant explanation, highlighting limitations in causal and inferential reasoning.

1 Introduction

Commonsense reasoning is a crucial aspect of Natural Language Processing (NLP) that enables models to understand and validate knowledge beyond explicit textual data. The motivation behind this research comes from the need to develop NLP models that can reason beyond surface-level text representations and apply real-world knowledge to language understanding tasks. Existing benchmarks, such as CommonGen (Lin et al., 2019), SemEval-2020 Task 4: Commonsense Validation and Explanation (Wang et al., 2020), CommonSenseQA 2.0 (Talmor et al., 2022), and COPEN (Peng et al., 2022), have highlighted various aspects of commonsense reasoning, including generative commonsense reasoning, multi-hop reasoning, and physical commonsense knowledge. However, these tasks still pose challenges in handling nuanced reasoning (El-Sayed and Pacholczyk, 2002), causal inference(Yao et al., 2021), and knowledge integration (Chen et al., 2020).

The SemEval-2020 Task 4: Commonsense Validation and Explanation (Wang et al., 2020) has served as a benchmark for evaluating various NLP models' capabilities in this domain. The task consistes of three sub-tasks, where in this research we will focus on the first two namely: Task A - Commonsense Validation: Determining whether a given statement aligns with commonsense knowledge, and Task B - Commonsense Explanation: Identifying the reasoning behind why a statement is against common sense. Table 1 provides examples on both tasks as they appear in the dataset.

This paper aims to explore how well large language models (LLMs) perform on commonsense reasoning tasks using zero-shot prompting. By evaluating multiple LLMs on SemEval-2020 Task 4, we investigate their ability to reason effectively without explicit fine-tuning. We present an overview of existing research, detail our methodology, and analyze experimental results to assess the strengths and limitations of current approaches.

2 Related Work

SemEval-2020 Task 4, which focuses on Commonsense Validation and Explanation, attracted considerable attention, with numerous teams participating in its three subtasks. This literature review highlights the best-performing models in Tasks A and B, showcasing their methodologies and contributions to the field.

CN-HIT-IT.NLP (Zhang et al., 2020) emerged as the leading model in Subtask A, employing a variant of K-BERT (Liu et al., 2019a) as its encoder. This model stands out for its integration of

Task	Example
Task A: Commonsense Validation	**Which statement is against common sense?** - **Statement 1:** He put a turkey into the fridge. (Correct) - **Statement 2:** He put an elephant into the fridge. (Against commonsense)
Task B: Commonsense Explanation	**Why is this statement against common sense?** **Statement:** He put an elephant into the fridge. - **A:** An elephant is much bigger than a fridge. (Correct) - **B:** Elephants are usually white while fridges are usually white. - **C:** An elephant cannot eat a fridge.

Table 1: Examples of Commonsense Validation and Explanation Tasks

knowledge graphs, specifically ConceptNet (Speer et al., 2017), which allows it to extract relevant triples that enhance the understanding of language representations. This approach underscores the importance of leveraging structured knowledge to improve commonsense reasoning capabilities.

In Subtask B, ECNU-SenseMaker (Zhao et al., 2020) achieved top performance by also utilizing K-BERT (Liu et al., 2019a). This model innovatively combines structured knowledge from ConceptNet (Speer et al., 2017) with unstructured text through a Knowledge-enhanced Graph Attention Network. This integration facilitates a deeper understanding of commonsense knowledge, demonstrating the effectiveness of combining different types of information to enhance model performance.

Another notable model, IIE-NLP-NUT (Xing et al., 2020), utilized RoBERTa as its encoder. This model's unique contribution lies in its second pretraining phase, which involved a textual corpus from the Open Mind Common Sense (OMCS) project (Singh et al., 2002). By exploring various prompt templates for input construction, this model illustrates the potential of tailored input strategies in improving commonsense validation tasks

Team Solomon (Srivastava et al., 2020) was ranked 5th and 4th in Subtasks A and B, respectively. Their approach, which also relied on RoBERTa, highlighted the capacity of large-scale pretrained language models to encapsulate commonsense knowledge effectively without external resources.

Across the two subtasks, the dominant trend was the use of large-scale pretrained language models such as K-BERT (Liu et al., 2019a), *RoBERTa* (Liu et al., 2019b), *BERT* (Devlin et al., 2018), and *GPT-2* (Radford et al., 2019), often fine-tuned with additional commonsense knowledge sources. Additionally, models incorporating external structured knowledge sources (e.g., ConceptNet) generally outperformed purely language-model-based approaches.

3 Methodology

Our study aims at evaluating the performance of multiple Large Language Models (LLMs) for commonsense validation and reasoning using zero-shot prompting. This approach leverages pre-trained LLMs without task-specific fine-tuning, relying solely on their inherent reasoning capabilities. For this purposes, we utilize the SemEval-2020 Task 4 dataset (Wang et al., 2020), which comprises labeled statements designed for commonsense validation and explanation tasks. To ensure a fair comparison between explicitly fine-tuned models and those evaluated solely with zero-shot prompting, we use only the test set for evaluation. The test set contains 1,000 entries for each task (Task A and Task B), providing a standardized benchmark for assessing model performance. The dataset is publicly available and can be accessed at [1].

As depicted in Figure 1, the methodology consists of the following key stages:

- **Pre-processing:** preparing the input test data templatic prompt to ensure compatibility with zero-shot prompting.
- **Model Calling:** Applying zero-shot prompting to multiple LLMs, including LLaMA3, Gemma2, and Mixtral to assess their commonsense validation and reasoning abilities.

[1] https://github.com/wangcunxiang/SemEval2020-Task4-Commonsense-Validation-and-Explanation

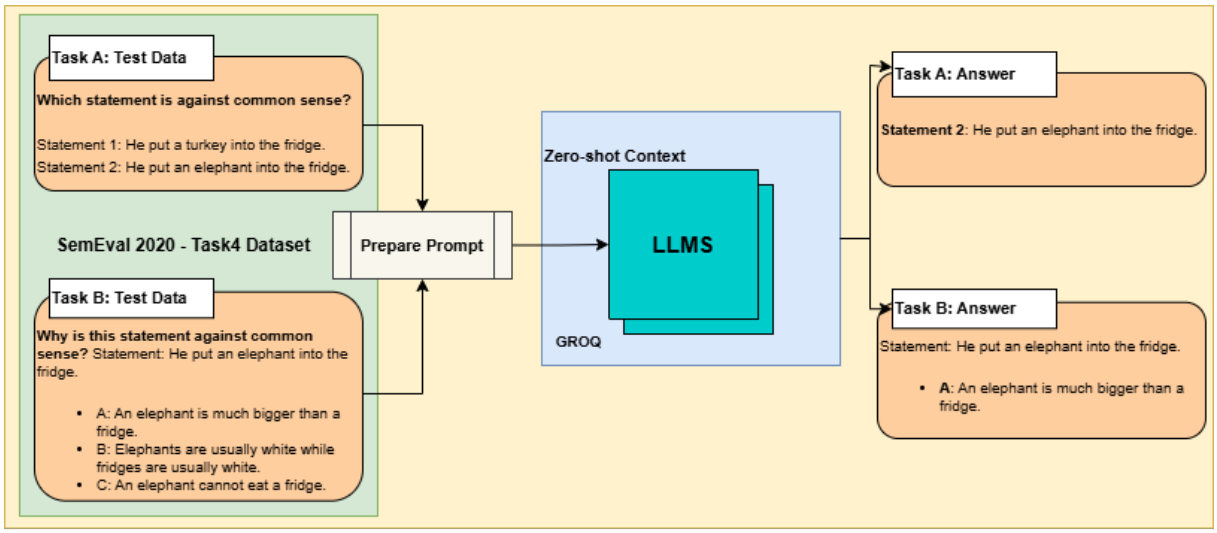

Figure 1: The architecture of the commonsense validation and reasoning with zero-shot prompting of LLMs.

LLMs are directly accessible through the GroqCloud [2] Models API endpoint using the model IDs

- **Performance Metrics:** Evaluating model outputs based on accuracy to quantify their effectiveness.
- **Comparative Analysis:** Benchmarking zero-shot LLMs performance against fine-tuned transformer models to examine the impact of training on commonsense validation and reasoning tasks.

4 Results and Discussion

Table 2 presents the performance of the models on the commonsense validation (Task A) and commonsense explanation (Task B) tasks from SemEval-2020 Task 4. The results for human performance and transformer-based models (CN-HIT-IT.NLP, ECNU-SenseMaker, IIE-NLP-NUT, and Solomon) are as reported in the original SemEval-2020 Task 4 paper (Wang et al., 2020). In contrast, the results for the LLMs (LLaMA3, Gemma2, and Mixtral) are obtained from our experiments with zero-shot prompting. Findings are reported in the following subsections.

[2] https://console.groq.com/docs/models

4.1 Performance Analysis

Among the models evaluated in this study, **L3-70B (LLaMA3-70B)** demonstrated the highest performance in Task A, scoring **98.4%**, with an evidence that large-scale LLMs can effectively validate commonsense knowledge with zero-shot prompting. However, its performance in Task B (**93.4%**) lags behind the transformer-based models reported as top 4 performing models in the Task paper. These models were explicitly fine-tuned for the task and some of them used external resources for the models training. This indicates that while LLMs are highly proficient in identifying implausible statements, they still struggle with selecting the most relevant explanation, demonstrating limitations in causal and inferential reasoning.

Similarly, the **G2-9B (Gemma2-9B)** model achieves strong performance in Task A (**97.9%**) but showing a more significant decline in Task B (**91.0%**). This further highlights the challenge of explanation selection, as these models may recognize implausibility without fully understanding the underlying causal mechanisms.

A size-dependent trend is observed in the LLaMA3 models. The smaller **L3-8B (LLaMA3-8B)** demonstrates significantly weaker performance than its larger version, with **84.4%** in Task A and **83.1%** in Task B. Finaly, the **M8x7B (Mixtral-8x7B)** model exhibited the weakest performance, with **66.0%** in Task A and **50.9%** in Task B. Its near-random performance in explanation selection

Model	Task A (Validation) (%)	Task B (Explanation) (%)
Human	99.1	97.8
CN-HIT-IT.NLP	97.0	**94.80**
ECNU-SenseMaker	96.7	95.0
IIE-NLP-NUT	96.4	94.3
Solomon	96.0	94.0
L3-70B (LLaMA3-70B)	**98.40**	93.40
G2-9B (Gemma2-9B)	97.90	91.00
L3-8B (LLaMA3-8B)	84.40	83.10
M8x7B (Mixtral-8x7B)	66.00	50.90

Table 2: Performance of different models on Task A (Commonsense Validation) and Task B (Commonsense Explanation) for English data. The models are: L3-70B (LLaMA3-70B), G2-9B (Gemma2-9B), L3-8B (LLaMA3-8B), and M8x7B (Mixtral-8x7B).

id	sent0	sent1	L3-70B	G2-9B	M8x7B	L3-8B
459	The dog pounced on the rabbit	The cat pounced on the rabbit	sent0	sent0	**Other**	sent0
737	She purchased four supermarket tickets.	She purchased four theater tickets.	sent1	sent1	sent1	sent1
174	Witches are not made of wood	Toads are not made of wood	sent0	sent0	sent0	sent0

Table 3: Sample of common misclassified instances for TaskA. Model abbreviations: L3-70B = LLaMA3-70B, G2-9B = Gemma2-9B, M8x7B = Mixtral-8x7B, L3-8B = LLaMA3-8B. Keep in mind that Task A is about identifying which statement is against common sense?

suggests that it struggles not only with causal inference but also with general commonsense understanding, likely due to limitations in its training data or architecture. It is important to note that this lower accuracy was not due to weak reasoning abilities but rather due to inconsistencies in the output format, where the model provided both classification and explanation instead of following the expected template for the output.

4.2 Implications for Zero-Shot Commonsense Reasoning

The results indicate that while LLMs often recognize implausible statements but fail to select the most relevant explanation, highlighting deficits in causal and inferential reasoning. This suggests that current zero-shot approaches may capture surface-level plausibility but lack deeper reasoning abilities necessary for explanation generation.

Furthermore, the comparison between pre-trained LLMs and task-specific models from SemEval-2020 Task 4 suggests that explicit fine-tuning on commonsense explanation data remains beneficial. While larger models such as **L3-70B** outperform fine-tuned models in validation, they do not surpass them in explanation selection, reinforcing the need for additional adaptation to improve causal reasoning.

4.3 Common Misclassification Patterns

An analysis of misclassified instances provides insights into the reasoning patterns of different models. In **Task A**, some models failed to differentiate between subtle variations in sentence structure. For example, the model incorrectly classified the following pair:

> *The dog pounced on the rabbit. The cat pounced on the rabbit.*

This type of error suggests that the models may rely on statistical patterns rather than deep semantic understanding.

In **Task B**, errors were primarily related to the selection of the most plausible explanation. A notable example is:

> **False Statement:** "There are four years each season."
> **Correct Explanation:** "A year can be divided into four seasons."

id	FalseSent.	OptionA	OptionB	OptionC	L3-70B	G2-9B	M8x7B	L3-8B
1388	Roberts' room is sleeping	A room cannot close his eyes, because he has no eyes.	Robert won't let the room sleep because he needs rest.	Robert can sleep in his room	A	A	A	A
1444	There are four years each season.	Different seasons have different temperatures.	A year can be divided into four seasons.	A season is shorter than a year.	C	C	**Other**	C
1172	People can need sleep.	Sleep is not a thing to have it granted.	Sleeping is nature for every living being.	Sleeping is an activity that every living thing does.	A	A	C	A

Table 4: Sample of common misclassified instances for TaskB. Model abbreviations: L3-70B = LLaMA3-70B, G2-9B = Gemma2-9B, M8x7B = Mixtral-8x7B, L3-8B = LLaMA3-8B. Task B is about selecting the reason for Why is this statement against common sense?

Some models selected incorrect explanations, indicating potential limitations in their ability to link cause-effect relationships effectively. It should be noted that sentence IDs **1388, 1444, and 1172** are not present in the common misclassified instances of Task A.

Despite the overall strong performance, the results also highlight challenges in certain reasoning aspects. The models demonstrated difficulty in selecting the most appropriate explanation for an implausible statement in Task B, even though they performed well in identifying implausible statements in Task A. This suggests that while the models recognize commonsense inconsistencies, they may struggle to justify their choices accurately. One possible explanation for this challenge is that Task B requires models to establish causal or inferential relationships between a false statement and its explanation. While Task A is a binary classification task requiring identification of implausible statements, Task B introduces additional complexity by demanding a deeper understanding of reasoning patterns and cause-effect relationships. Selecting the correct explanation requires not only recognizing a logical inconsistency but also evaluating multiple plausible justifications and determining which one best aligns with human commonsense knowledge. This suggests that current LLMs, despite their powerful language modeling capabilities, may still struggle with selecting the most contextually relevant explanation among multiple plausible options, as this task requires a nuanced understanding of real-world implications and reasoning structures (Mondorf and Plank, 2024).

Additionally, the low measured performance of **Mixtral-8x7B** can be attributed to output inconsistencies. The model frequently produced both an answer and an explanation, which deviated from the required response format. This indicates that we cannot rely on the achieved results for this model to evaluate its performance on both tasks. More post-processing steps are required to ensure consistent output formatting when evaluating model performance.

4.4 Conclusion

This study demonstrates that large-scale LLMs, particularly **LLaMA3-70B** and **Gemma2-9B**, exhibit strong commonsense reasoning capabilities even in a zero-shot setting. These models outperform state-of-the-art fine-tuned transformer-based models, indicating that LLMs can generalize well across commonsense validation tasks. However, challenges remain in explanation selection and maintaining consistent output formats. Future research may include exploring Commonsense knowledge-graph

LLMs (Li et al., 2022; Zhao et al., 2024; Toroghi et al., 2024), in addition to fine-tuning strategies, retrieval-augmented approaches, and structured prompting techniques to enhance the inferential reasoning capabilities of LLMs in zero-shot settings.

References

Xiaojun Chen, Shengbin Jia, and Yang Xiang. 2020. A review: Knowledge reasoning over knowledge graph. *Expert systems with applications*, 141:112948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*.

Mazen El-Sayed and Daniel Pacholczyk. 2002. A qualitative reasoning with nuanced information. In *Logics in Artificial Intelligence*, pages 283–295, Berlin, Heidelberg. Springer Berlin Heidelberg.

Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A systematic investigation of commonsense knowledge in large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2019. Commongen: A constrained text generation challenge for generative commonsense reasoning.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019a. K-bert: Enabling language representation with knowledge graph. In *arXiv preprint arXiv:1909.07606*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey. *arXiv preprint arXiv:2404.01869*.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. Copen: Probing conceptual knowledge in pre-trained language models. *arXiv preprint arXiv:2211.04079*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *Technical Report*.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237, Berlin, Heidelberg. Springer Berlin Heidelberg.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Vertika Srivastava, Sudeep Kumar Sahoo, Yeon Hyang Kim, Rohit R.r, Mayank Raj, and Ajay Jaiswal. 2020. Team Solomon at SemEval-2020 task 4: Be reasonable: Exploiting large-scale language models for commonsense reasoning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 585–593, Barcelona (online). International Committee for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. Commonsenseqa 2.0: Exposing the limits of ai through gamification.

Armin Toroghi, Willis Guo, Mohammad Mahdi Abdollah Pour, and Scott Sanner. 2024. Right for right reasons: Large language models for verifiable commonsense knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6601–6633, Miami, Florida, USA. Association for Computational Linguistics.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. Semeval-2020 task 4: Commonsense validation and explanation. In *Proceedings of The 14th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Luxi Xing, Yuqiang Xie, Yue Hu, and Wei Peng. 2020. IIE-NLP-NUT at SemEval-2020 task 4: Guiding PLM with prompt template reconstruction strategy for ComVE. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 346–353, Barcelona (online). International Committee for Computational Linguistics.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46.

Yice Zhang, Jiaxuan Lin, Yang Fan, Peng Jin, Yuanchao Liu, and Bingquan Liu. 2020. Cn-hit-it.nlp at semeval-2020 task 4: Enhanced language representation with multiple knowledge triples. In *Proceedings of The 14th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Qian Zhao, Siyu Tao, Jie Zhou, Linlin Wang, and Xin Lin. 2020. Ecnu-sensemaker at semeval-2020 task 4: Leveraging heterogeneous knowledge resources for commonsense validation and explanation. In *Proceedings of The 14th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2024. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36.