Little Minds, Big Laughs: Supercharging Small LLMs for Humor Detection

Andres Cubeddu, Muhammad Jee, Joseph Greene, Arno Motulsky

Abstract

Humor detection is a challenging task for large language models (LLMs) due to the nuanced and context-dependent nature of humor. This paper explores methods to enhance humor detection in smaller-sized LLMs, such as Mistral 7B and Llama 3 8B, by incorporating context. We experiment with three approaches to providing context across two distinct humor tasks: N-Sentence context, Retrieval-Augmented Generation (RAG) context, and External Context. We find that additional context does not consistently improve model performance. Rather, our findings suggest that context often confuses the model. These results have significant implications for the development of efficient, low-cost, and low-latency AI systems capable of nuanced understanding. Our research can help harness these models to their full extent by using non-traditional methods to improve performance.

KEYWORDS

Humor Detection; Large Language Models; Contextual AI; Low-Resource AI Models; Retrieval Augmented Generation

1. Introduction

Humor is a fundamental element of human interaction that does more than just make people laugh. It extends beyond cultural confines and plays an integral role in social interaction, fostering connections and deepening trust [1]. Humans possess an innate ability to grasp, interpret, and generate humor with ease in their everyday interactions. However, the discomfort of missing humor or mistaking something serious as humorous shows how important accurate humor detection is to people. Explaining why something is humorous is not always an easy task [2]. The subjective nature of evaluating humor demands a nuanced understanding of language, context, and socio-cultural factors [3]. Humor is intricately linked to culture, and interpreting humorous communication demands an awareness of diverse cultural presuppositions [3-4]. This subjectivity is further complicated by individual differences in humor appreciation, which can be influenced by personal experiences, mood, and cognitive styles [5]. The complexity of humor detection makes computational humor detection a uniquely difficult task, requiring models to grasp subtleties that are often challenging even for humans to articulate.

Advancements in computational humor detection would significantly enhance Natural Language Processing (NLP) capabilities, paving the way for more sophisticated and intelligent AI systems. Previous work has utilized Large Language Models (LLMs) to detect and explain humor, but these models are often computationally expensive, relying on very large LLMs [6-8]. Our work aims to improve the performance of smaller models on humor detection tasks by providing context, thereby avoiding significant increases in computational power. Context enriches the comedic experience. It transforms what might otherwise be a mundane statement into a source of amusement. Jokes often rely on background information and situational nuances to generate laughter, and without the appropriate context, a joke may fall flat. Specifically in sitcoms, the humor is often embedded in the intricacies of the characters, the dynamics of the scene, and the previous dialogues. Understanding the characters' personalities, their relationships, and the preceding events is often essential for fully appreciating the humor in a given moment.

In this paper, we investigate how different methods of incorporating additional context can enhance a model's performance in humor detection tasks. We specifically examine three approaches: N-Sentence context, which involves using a n-number of preceding sentences to provide immediate context; Retrieval-Augmented Generation (RAG)-based context, which gathers relevant information retrieved from a dataset; and External Context, which uses information from external sources to provide a broader background. By exploring these diverse methods, we aim to identify how context can improve the accuracy and effectiveness of humor detection models.

The field of natural language processing is witnessing the rapid emergence of increasingly larger and better-performing LLMS. While these models achieve impressive results across various tasks, including humor detection, their enhanced performance often comes at the cost of significantly higher computational resources. Researchers facing poor model performance with smaller models frequently resort to the easier route of improving performance by implementing more computationally expensive models. This paper aims to address this issue by exploring the potential of leveraging context to improve humor detection tasks. Our research in improving smaller language models like Mistral 7B or Llama 3 8B for humor detection holds significant potential across diverse sectors, especially in scenarios where low latency and reduced dependency on internet connectivity are crucial. In developing countries, this research could aid in the development of low-cost communication technologies like personal chatbot AI systems that operate efficiently through SMS. Our work can help make AI-based humor and interactive conversational engagement accessible even in regions with limited internet infrastructure. This also opens up new avenues for education, information dissemination, and entertainment. For media companies in these regions, these models can present an opportunity to produce localized, humorous content more effectively. We hope to build upon the limited literature in this space an area that has seen less focus compared to advancements in larger models.

1.1. Related Work

Previous work in computational humor has primarily focused on using computationally intensive methods, including very large and multimodal models. We use the following works' data and are guided by some of their results.

Hessel et al. (2023) explored whether large language models (LLMS) can effectively understand humor through several tasks, using over a decade's worth of data from New Yorker cartoon caption contests [9]. The tasks include caption-cartoon matching, caption ranking, and caption humor explanation. The researchers contribute annotations to the dataset for each cartoon describing the situation, detailing what might be off, and explaining the joke. They found that humans significantly outperform all models on all tasks but that larger models, such as GPT-4, performed better than smaller ones, like T5-Large and T5-11B.

Patro et al. (2021) examine whether providing visual and textual context can improve model performance on classifying sets of dialogue from the sitcom The Big Bang Theory as humorous or non-humorous. The researchers create an annotated dataset of dialogues, including the characters speaker and listening and a short description of the scene. While Patro et al. employ multimodal models in their research, their best text-only result was an F1 score of 79.96% (for the humor class) and 67.98% accuracy overall. We make use of Patro et al.'s dataset and methods to create our own modified versions [10].

1.2. Contributions

In this paper, we present several significant contributions to the field of humor classification using smaller-sized LLMs. The main contributions of our work are: (1) proposing a novel methodology for investigating the effects of context on smaller-sized LLMs, which includes N-Sentence context, Retrieval-Augmented Generation (RAG) context, and External Context; (2) designing and executing experiments across two popular smaller-sized LLMs and uncovering insights into limitations; (3) creating and releasing an updated dataset of "The Big Bang Theory" scripts by uniquely tagging each dialogue line with a humor label.

2. Methodology

2.1. Datasets

Our study incorporates two datasets to develop a holistic understanding of humor in smaller-sized LLMS.

Multimodal Humor Dataset (D1)

This dataset consists of scripted dialogue from 'The Big Bang Theory' covering the first five seasons, providing a total of 25,701 dialogue instances. Each dialogue is annotated as either humorous or non-humorous based on the subsequent audience laugh tracks, offering a direct linkage to perceived humor. The dialogues include detailed information about the speaker, the recipient(s), the scene, and the timing of the dialogue. This dataset is valuable for exploring situational humor, where temporal and conversational dynamics play a role in the humor's emergence.

We modified and released this dataset with several significant changes. First, we transformed each data point into a unique dialogue line instead of a set of dialogues. In the original dataset, individual lines were not tagged as humorous or non-humorous; instead, sets of lines included timestamp annotations for laugh tracks. We revised the dataset to include the following information for each line: "scene," "recipients," "speaker," "dialogue," and "isHumor." The "isHumor" variable is a binary indicator that denotes whether a humor timestamp immediately followed the given line. This modification allowed us to exclude instances of visual humor. For example, if a laugh track occurred at the start of a scene with no dialogue, it would be excluded from our dataset since the humor stemmed from visual elements, not the dialogue.

New Yorker Cartoon Caption Dataset (D2)

Compiled from 14 years of The New Yorker's weekly cartoon caption contests, this dataset includes over 700 cartoons and thousands of associated captions. Each contest entry comprises the cartoon itself, all submitted captions, the top three finalists selected by the editors, and occasionally, crowd-sourced quality ratings for each caption. This dataset focuses on conceptual humor, where humor arises from understanding and manipulating concepts and the relationships between them.

2.2. Models & Prompts Design

We conduct our experiments using pre-trained Language Models based on the transformer architecture. We use the Hugging Face and OpenAI API to query the following text models: Mistral-7B-Instruct, Llama-3-8B-Instruct, Gpt3.5-turbo, and Gpt4. Since this ordering reflects increasing computational power and claimed capability, we expect that their performance would show increasing fidelity.

These models process a text prompt and generate a randomized completion, generating text that would logically follow the prompt based on its training data. Liu et al. review various approaches to crafting prompts to utilize these models for text generation [11]. OpenAI's guidelines highlight that effective prompts should include a system message with a clear role and instructions [7]. Research also demonstrates possible utility in providing the model with a few illustrative examples, a technique known as few-shot prompting. However, LLMs can be highly sensitive to the selection and order of these examples [12]. For our experiments, we use a standardized prompt structure with slight adjustments according to the experiment and feed it into the LLMs.

2.3. Contextual Memory

In our experiments, we use a variety of strategies to enhance the models' ability to understand and classify humor by enriching their inputs with different layers of context. In our experiments, we carefully designed our prompts to guide the models effectively. Each prompt consists of a system message containing task instructions and a user message with context and target dialogue.

N-Sentence Context

This approach utilizes the preceding 'n' sentences from a dialogue as a context for the model. The intent is to see if providing prior conversational history enhances the model's ability to correctly classify humor based on the progression and buildup of the current conversation.

We converted our raw dialogue and metadata into a more natural format, such as by including `[speaker] says "[dialogue]" to [recipients]`. We also include scene descriptions, such as "In the cafeteria," and note in the prompt when the scene changes between lines. We distinguish between the context and target lines in the prompt with either "Context:" or "Use these lines as context:". We employ a chat template, including one message from the system role containing instructions for the task and one user message with the context, the target, and the question reiterated. Including the question at both the beginning and end seemed to coax the models into responding only with the acceptable answers ("0", "1").

System Message:
You are a humor classification model. Determine whether the target dialogue is humorous using the previous dialogue as context.

User Message:
Context: [N-Sentence Context]
Target: [Target Dialogue]

Is the target dialogue humorous (return 1) or not humorous (return 0)?

Figure 1. Prompt template used for humor classification with N-Sentence Context.
Ok

RAG-Based Context

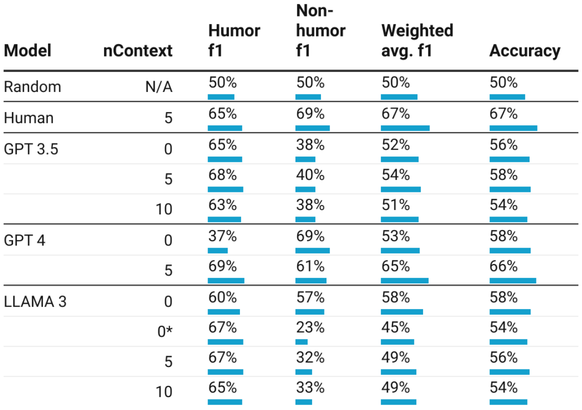
Retrieval-Augmented Generation (RAG) involves dynamically pulling in relevant information (dialogues/scenes) from the same episode or the same season. This method tests whether broader contextual understanding from the series improves humor classification compared to isolated dialogue snippets. We employed the LangChain framework and FAISS (Facebook AI Similarity Search) for this task. For each episode, we created embeddings using the Hugging Face sentence-transformers model. These embeddings were stored in a vector store, which allowed efficient retrieval of relevant chunks during inference. We use different values of k (2, 5, and 10) to explore the impact of the number of retrieved chunks on the model's performance. The retrieval process excludes the exact dialogue to avoid redundancy and focuses on providing context that enhances understanding. The prompt included retrieved context chunks formatted into natural dialogue. Scene descriptions and changes between lines are included to provide comprehensive contextual information.

External Context

We provide the model supplementary information beyond the immediate text or context. For D1, we incorporate a synopsis of the show along with detailed data about the show's main characters. This may help the model grasp nuances of character-driven and situational humor by connecting dialogues to broader narrative elements. For D2, we provided text chunks from relevant Wikipedia articles to provide the models with additional cultural, historical, and conceptual information. This may help concentrate the model's attention on immediate, relevant information.

3. Results

N-Sentence Context



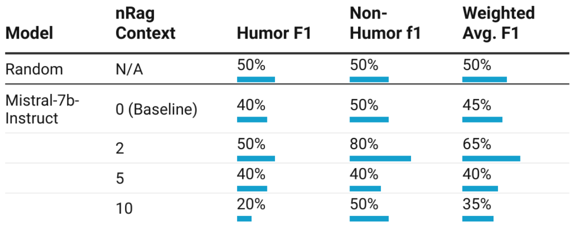
Model	nContext	Humor f1	Non-humor f1	Weighted avg. f1	Accuracy
Random	N/A	50%	50%	50%	50%
Human	5	65%	69%	67%	67%
GPT 3.5	0	65%	38%	52%	56%
	5	68%	40%	54%	58%
	10	63%	38%	51%	54%
GPT 4	0	37%	69%	53%	58%
	5	69%	61%	65%	66%
LLAMA 3	0	60%	57%	58%	58%
	0*	67%	23%	45%	54%
	5	67%	32%	49%	56%
	10	65%	33%	49%	54%

Table 1. n-Sentence Context results.
**with scene descriptions.*

We tested each model using 200 randomly sampled instances from the first five seasons of The Big Bang Theory. The set of samples was consistent across each test (for each context length) and balanced between the humor and non-humor classes, with 100 samples for each.

The nContext experiment results show that on a (relatively) small model like Llama 3 8B, adding context yielded poorer overall performance. While the models achieved higher f1 scores on the humor class when some context was infused, this came with significantly poorer performance on the non-humor class. Larger models like GPT 3.5 and 4, however, saw improved performance in all areas when some context was added, except for GPT 4's slightly lower non-humor f1 score compared to its baseline. In all tests, 10-sentence context yielded poorer performance in all areas than 5-sentence context. Moreover, we see from our human baseline that humor classification tasks are inherently quite difficult.

RAG-Based Context



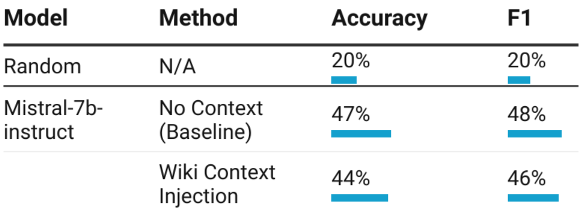
Model	nRag Context	Humor F1	Non-Humor f1	Weighted Avg. F1
Random	N/A	50%	50%	50%
Mistral-7b-Instruct	0 (Baseline)	40%	50%	45%
	2	50%	80%	65%
	5	40%	40%	40%
	10	20%	50%	35%

Table 2. RAG-Based Context results.

The RAG-based context experiment reveals that the Mistral-7b-Instruct model's humor classification performance significantly improves with a brief, relevant context but deteriorates with excessive information. Without context, the model's performance was modest, slightly favoring non-humorous detection. Introducing a 2-dialogue context boosted the weighted Humor F1 score from 0.45 to 0.65. However, increasing the context to 5 and 10 dialogues resulted in a decline in performance.

Comparing the baselines of this experiment to the N-Sentence, we saw a 0.15 decline in the weighted F1 score for Retrieval Augmented Generation. However, when given two pieces of retrieved context, Mistral performs about as well as our human tester and GPT 4. However, similar to the results of the N-Sentence context experiment, giving the model more context yields poorer performance.

New Yorker Matching Task



Model	Method	Accuracy	F1
Random	N/A	20%	20%
Mistral-7b-instruct	No Context (Baseline)	47%	48%
	Wiki Context Injection	44%	46%

Table 3. New Yorker Matching Task results.

Mistral performs better than random guessing (20%) at both the baseline and with Wikipedia context. However, the model performs worse with Wikipedia context (44%) than without (47%).

4. Discussion

Although we hoped to discover that alternative methods from increasing model size–such as context injection–improved the performance of small LLMs on humor tasks across two distinct tasks, they generally did not. In most cases, additional context injection actually produced worse results than the baseline. Only in the specific case of additional sitcom context selected via two-sentence Retrieval Augmented Generation from the same episode did we observe performance improvements above the baseline.

Notably, for the cartoon caption matching task, the baseline included detailed descriptive information, so context injections included *additional* context that might aid humor understanding outside of the joke itself. In the case of the sitcom task, the 0-sentence context baseline included no scene descriptions or additional context beyond the target sentence itself. While the inclusion of previous sentences as context might, in theory, serve to provide a greater understanding of the scene and, as a result, the target sentence, in practice, the additional context did not improve performance on the task as a whole. More specifically, adding context to the Llama-3 8B model prompts, ranging from scene descriptions to 5 and 10 sentences of directly previous dialogue, significantly decreased performance on true non-humorous target lines while only slightly increasing performance on true humorous lines.

We suspect that the large degradation of performance on non-humorous examples that results from additional context is associated with polluting or clouding the model's objective. That is, the inclusion of additional context beyond the target line adds content to the prompt that may itself be humorous, likely influencing the model as the model conflates the context with the target.

Reference results from GPT-4 and GPT-3.5, much much larger models, performance on the same task indicate that not only delineating the context from the target but also ‘understanding’ and using the additional context effectively to inform judgment of the target is a complex and difficult task–one that benefits from the complexity afforded by larger models. It is remarkable that even a model as large as GPT-3.5 may not be afforded this ability by its model size: the effects of additional context on the performance of GPT-3.5 are more similar to Llama-3 8B than they are to GPT-4. Yet, it is clear from the performance improvement of GPT-4 associated with 5-sentence context that additional context does offer the potential for improved performance under favorable conditions.

Similar findings from the cartoon caption matching task support our belief about model distraction and the elusiveness of complex understanding. In the cartoon caption matching task, the baseline included detailed descriptive information that was required to effectively translate a cartoon image into text, so context injections in this task were quite additional. Similarly, introducing additional information to the prompt effectively reduced the importance of the core information, watering it down or adding more distracting information. For example, processing the Wikipedia article text for ‘conference room’ may lead the model astray more than it helps the model better understand a nuanced joke that plays on conference room culture and experience. A future experiment to test this hypothesis more comprehensively might involve examining attention patterns via layer probing to gain a better understanding of the models’ comprehension of additional context.

Importantly, selecting context for the sitcom task that was semantically similar (via RAG) rather than in close proximity (N-Sentence) offered a huge performance improvement over the baseline for both true humorous and non-humorous targets: 30pts and 25pts respectively. We will note that we used Mistral 7B for the RAG sitcom task as well as the cartoon caption matching task, but we believe that the comparisons to the baselines provided for both models serve as ample context for each model.

We believe that selecting context based on semantic similarity rather than positional relevance serves to mitigate the potentially distracting or diluting effects that context can have, making clear the importance of semantically relevant context. This notion is furthered by the comparative performances of 2-sentence, 5-sentence, and 10-sentence semantically similar context. That is, while the inclusion of two sentences of similar context significantly improves performance compared to the baseline, the inclusion of more sentences of similar context results in equal or worse performance compared to the baseline. Thus, more context oftentimes does not improve performance, but small amounts of carefully selected, semantically similar context does.

We expected this task to be difficult to perform with small LLMs based on the literature. Wei et al. find that Chain of Thought prompting, which invokes a similarly complex model understanding, only produces significant improvements in model performance with large LLMs, even leading to worse performance than the baseline in small models [13]. This notion of unique ability produced by scaling up model sizes is also described by Wei et al. in their discussion of emergent properties of large language models [14]. Similarly, we believe that the ability of models to be able to separate context from target as part of forming a complex and nuanced notion of understanding that extends far beyond simple neural connections between words and phrases is also produced by larger model sizes, in line with the notion of emergent abilities. Importantly, we find that specific techniques of including additional context, namely Retrieval Augmented Generation, help mitigate the limits of small models and provide improved performance without increasing computational power. The Mistral 7B release paper argues for an emphasis on three dimensions of model improvement (training cost, inference cost, and model capabilities) rather than the traditional emphasis on just two (training cost, inference cost) (5). Our finding that prompting techniques can significantly improve small model performance without increasing computational power invites the possibility of a fourth, adjacent dimension: prompt content quality in deployment.

5. Future Work

In our follow-up research, we will enhance our experiments by incorporating advanced frameworks to refine how LLMs make decisions and reason through humor-related tasks. Using a decision-tree framework, we aim to improve the logic flow in the LLMs prompting mechanism. Additionally, we will implement multi-step reasoning, enabling the models to break down complex classification problems into intermediate steps, further enhancing their reasoning capabilities.

Moreover, to test our hypothesis that additional context distracts smaller models from the core task, we propose an experiment using layer probing to examine attention patterns. We will utilize a dataset of jokes that rely on specific contexts, preparing two versions of the input: one with only the joke and another with the joke and related context. By selecting small language models and analyzing attention distributions across layers, we aim to identify shifts in focus caused by the additional context. This will help us understand how added information impacts the models' ability to concentrate on the core humor-related task.

Additionally, in our experiments using sitcom data, we assume our algorithm for identifying humorous lines based on laugh track timing performs sufficiently well on this classification task. However, there may be misclassified instances. Future work will also explore whether including humorous context negatively impacts performance on the non-humor class, which could explain part of the models' decreased performance when provided with additional context.

6. Conclusion

Our study into improving humor detection in smaller language models reveals that adding context does not consistently improve performance and can sometimes hinder it. Despite our initial hypothesis, both N-Sentence and Retrieval-Augmented Generation (RAG) contexts often resulted in decreased accuracy. Notably, only specific cases of context, such as a two-sentence RAG from the same episode in sitcom tasks, led to improvements.

The results suggest that simply providing more context is insufficient; the nature and quality of the context are crucial. Semantically relevant context showed promise, particularly when limited to a small subset of context. This insight opens new avenues for refining context selection techniques and developing smarter context integration methods.

Our work highlights the potential and limitations of using context to improve humor detection in low-resource models. An implication of our findings is the potential for developing more accessible AI technologies. By refining the techniques for context integration in smaller-sized LLMs, we can create AI systems that are not only cost-effective but also capable of performing complex tasks like humor detection in environments with limited computational resources. This can democratize access to advanced AI capabilities, making them available in education, healthcare, and entertainment sectors worldwide.

GitHub Repository

https://github.com/adc257/info4940-sitcom

References

[1] Fritz, H. L. (2020). Why are humor styles associated with well-being, and does social competence matter? Examining relations to psychological and physical well-being, reappraisal, and social support. Personality and individual differences, 154, 109641.

[2] Kianbakht, S. (2020). Towards a comprehensive theory of culturally constructed humour. European Journal of Humour Research, 8(2), 1-24. https://doi.org/10.7592/ejhr2020.8.2.kianbakht

[3] Heydon, G. and Kianbakht, S. (2020). Applying cultural linguistics to translation studies: a new model for humour translation. International Journal of Comparative Literature and Translation Studies, 8(3), 1. https://doi.org/10.7575/aiac.ijclts.v.8n.3p.1

[4] Mitchell, H. H., Graesser, A. C., & Louwerse, M. M. (2010). The Effect of Context on Humor: A Constraint-Based Model of Comprehending Verbal Jokes. Discourse Processes, 47(2), 104–129. https://doi.org/10.1080/01638530902959893

[5] Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2017). Ha ha? Assessing individual differences in humor production ability. Psychology of Aesthetics, Creativity, and the Arts, 11(2), 231.

[6] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

[7] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), Advances in Neural Information Processing Systems, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020a.

[8] Kenneth, M. O., Khosmood, F., & Edalat, A. (2024, January 30). Systematic Literature Review: Computational Approaches for Humour Style Classification. arXiv.org. https://arxiv.org/abs/2402.01759

[9] Hessel, Jack & Marasovic, Ana & Hwang, Jena & Lee, Lillian & Da, Jeff & Zellers, Rowan & Mankoff, Robert & Yejin, Choi. (2023). Do Androids Laugh at Electric Sheep? Humor “Understanding” Benchmarks from The New Yorker Caption Contest. 688-714. 10.18653/v1/2023.acl-long.41.

[10] Patro, B. N., Lunayach, M., Srivastava, D., Sarvesh, S., Singh, H., & Namboodiri, V. P. (2021). Multimodal humor dataset: Predicting laughter tracks for sitcoms. In Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021 (pp. 576-585). Article 9423266 (2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021). IEEE.

[11] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021. URL https://arxiv.org/abs/2107.13586.

[12] Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. CoRR, abs/2104.08786, 2021. URL https://arxiv.org/abs/2104.08786.

[13] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. ICLR.

[14] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. Transactions on Machine Learning Research.