
OverPrompt: Enhancing ChatGPT through Efficient In-Context Learning

Jiazheng Li^{*1}, Runcong Zhao^{*1}, Yongxin Yang³, Yulan He^{1,2}, and Lin Gui¹

¹Department of Informatics, King’s College London, UK

²The Alan Turing Institute, UK

³Queen Mary University of London, UK

Abstract

The remarkable performance of pre-trained large language models has revolutionised various natural language processing applications. Due to huge parameter sizes and extensive running costs, companies or organisations tend to transfer the models to the target task by zero-shot prompting techniques. However, the prohibitive costs of tokens and time have hindered their adoption in applications. We propose OverPrompt, leveraging the in-context learning capability of LLMs to handle multiple task inputs, thereby reducing token and time costs. This approach could potentially improve task performance during API queries due to better conditional distribution mapping. Evaluated across diverse classification datasets, our experiments show that OverPrompt can achieve cost-efficient zero-shot classification without causing significant detriment to task performance, and in some cases, even improving it. An ablation study conducted on various LLMs, along with an investigation into the robustness of our prompting strategy to different input ordering, offers valuable insights into the broader applicability of our method across diverse tasks. These findings also suggest a more seamless integration of our method with LLMs through an API.

1 Introduction

Large Language Models (LLMs), such as ChatGPT (Ouyang et al., 2022), have shown emergent abilities on various Natural Language Processing (NLP) tasks (Wei et al., 2022). In-context learning (ICL) approaches, including zero-shot or few-shot prompting strategies (Kojima et al., 2022; Brown et al., 2020; Kaplan et al., 2020), offering computational efficiency and considerable performance gains without the need of fine-tuning LLMs. Various research has been carried out to improve the performance of zero/few-shot learning, such as searching for better few-shot examples (Zhang et al., 2023; Wang et al., 2023b) or finding appropriate prompts (Wang et al., 2023c; White et al., 2023).

However, cost-efficient prompting strategies are under-explored. Given the immense parameter size of LLMs, deploying them locally for certain industrial applications becomes impractical. Consequently, the substantial time and token costs associated with accessing these models through APIs present a significant challenge when adopting these models in production environments. While few-shot prompting entails the inclusion of demonstration examples, thereby raising the token cost of API queries, in most zero-shot prompting cases (Mialon et al., 2023; White et al., 2023), the bulk of the input content is allocated to the task description, leaving only a portion for the task input. The repetition of the task description can result in a substantial cumulative cost for each individual query. Hence, it becomes imperative to reduce the token and time costs associated with utilising these LLMs.

^{*}Equal contribution. Email: {jiazheng.li, runcong.zhao}@kcl.ac.uk

To address this issue, we propose OverPrompt, a zero-shot prompting strategy designed to process multiple instances simultaneously in a single query to enhance efficiency. Leveraging the emergent capability of LLMs, known as ICL, we analyse our prompting strategy within a Bayesian inference framework. Theoretically, our prompting strategy ensures better approximation of input task distributions by incorporating additional data and mitigating format errors. We also empirically show that our designed sampling and formatting framework enhances performance. In order to understand the overall impact of OverPrompt on query efficiency, we evaluate OverPrompt across ten different text classification datasets. Our experiments reveal that OverPrompt reduces both token and time costs, while leveraging the in-context learning capabilities of LLMs to produce improved conditional distributions for tasks when additional instances are provided. We also modify the output formatting to address performance degradation and reduce errors. Performance enhancements are observed when contextual information supplements the model’s decision-making process. This is particularly useful in tasks such as fact-checking, where extra evidence or logical deductions can be provided, and sentiment analysis, where well-defined category boundaries can be established through comparison. Nevertheless, tasks like sentence entailment may not gain any advantage from such context input¹.

2 OverPrompt

In this section, we introduce OverPrompt, a zero-shot classification strategy that utilizes ChatGPT’s ICL ability for more efficient zero-shot classification, reducing token consumption and time costs.

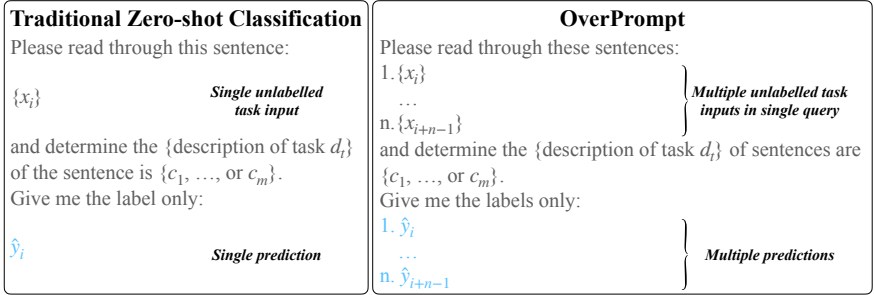


Figure 1: This illustration highlights the difference between traditional zero-shot classification prompting strategy and our OverPrompt. Deploying LLMs requires significant computational resources, so abandoning API queries is not practical. Our OverPrompt strategy prioritizes cost efficiency while maintaining task performance. It achieves this by reusing task descriptions and batch-processing task inputs, which reduces token usage and the number of API calls.

In a text classification setup, let $X = \{x_i\}_{i=1}^N$ be the text input, and $Y = \{y_i\}_{i=1}^N$ be the associated label set for the given input. The label y_i belongs to a category set $C = \{c_i\}_{i=1}^m$. For each prompt, the dataset element x_i is incorporated into the task-related description d_t , which introduces the target category set for label prediction. For instance, in the SST-2 dataset, the task description is "Sentiment Analysis", and the label set C comprises "positive" and "negative". We use \hat{y}_i to denote the predicted label for a given task input x_i and **highlight the prediction in blue**.

Figure 1 shows a traditional zero-shot classification prompt template on the left-hand side (Wang et al., 2023a). However, this template has two main limitations when it is applied: Firstly, it requires time-consuming iterations to process each task input, which can be expensive considering the delay of internet connections. Secondly, the current zero-shot classification paradigm only considers the task description d_t while ignoring connections between the task inputs. Previous research has not explored how multiple unlabelled task inputs might aid LLMs in using their ICL capability to determine a suitable task-related batch grouping, which might lead to a more accurate output generation

Building on the work of Xie et al. (2022), who used Bayesian inference to interpret the ICL capability of LLMs, we assume an LLM can perfectly fit the pre-training distribution p_θ with sufficient data, i.e., $p_{LLM} = p_\theta$. The crux is to extract the hidden concept θ^* from the given prompt (d_t, X_j) and use it to derive the conditional distribution $p_{\theta^*}(y_i|x_i)$, where $x_i \in X_j$. Our aim is to investigate whether $\text{argmax}_{y_i} p_\theta(y_i|d_t, X_j \setminus x_i, x_i) \rightarrow \text{argmax}_{y_i} p_{\theta^*}(y_i|x_i)$ as n increase. In other words, we

¹Our code is available at <https://github.com/lijiazheng99/OverPrompt>.

want to understand whether the LLM becomes more effective at making a more accurate predictions on the testing examples when provided with a larger set of samples X_j to aid in the inference of the prompt concept θ^* . This convergence holds under a distinguishability condition wherein θ^* is distinguishable:

$$\forall \theta \in \Theta, \theta \neq \theta^*, \epsilon_{start}^\theta + \epsilon_{delim}^\theta \leq \sum_{j=1}^k KL_j(\theta^* || \theta) \quad (1)$$

where the k is number of task inputs, ϵ_{start}^θ and ϵ_{delim}^θ represent errors of mismatches between the prompt and pre-training distributions and the delimiter token for each task input, which is bounded by the KL-divergence of corresponding difference of prompt and pre-trained distributions.

Obviously, the right-hand-side of Eq.1 increases with a larger number of examples k , improving distinguishability (Xie et al., 2022). In other words, the context of task inputs, beyond just the input-output mapping, can be valuable for ICL. Therefore, we introduce OverPrompt, a zero-shot classification prompt strategy that utilises LLMs’ emergent ICL capability by increasing the number of task inputs included in the prompt to n , as shown on the right-hand-side of the Figure 1.

Our proposed strategy involves finding a partition of input text set X where $\bigcup_{j=1} X_j = X$ and $\bigcap_{j=1} X_j = \phi$. The predicted labels can be obtained by LLM($\hat{Y}_j | d_t, X_j$), where \hat{Y}_j represents the predicted labels corresponding with input texts X_j . This strategy enables LLMs to handle multiple inputs simultaneously. Our experiments in §3.1.1 demonstrate that this approach significantly reduces query and lag time by reducing the number of API requests. Additionally, OverPrompt reduces the number of input tokens due to the shared prompt information base (e.g., task description d_t , the label set C), resulting in lower token usage costs.

Besides, comprehensively, prompt grouping can provide task-specific hints. This is because task-specific tokens usually appear more often in the inputs than general corpora. Our proposed strategy amplifies these informative words by grouping together semantically similar instances (grp), which is able to help concentrate $p(\theta | d_t, X_j)$ on the prompt concept with more examples. This, in turn, facilitates “locating” the concept θ^* . We also propose additional grouping strategies for ablation studies: (a) mixing these topics with random samples (mix), and (b) filtering mix to keep only topic-specific instances (fil). We provide detailed comparisons in §3.1.2.

OverPrompt can improve the performance of semantic meaning focused classification tasks like sentiment analysis, or fact-checking. However, as demonstrated in section A.2, adding more training instances does not always lead to better results for inferencing-related tasks. The i.i.d nature of training examples can cause unnatural transitions when randomly concatenated, which introduces noise and mismatches between the pre-training and prompt distributions. This can have a negative impact on performance, as observed in natural language inference tasks.

Output Formatting: Mitigating Performance Degradation and Errors As the number of outputs increases, we may encounter issues with inconsistencies in output formatting, resulting in a mismatch error. While most inconsistencies can be resolved using rule-based post-processing methods, mismatches where the number of outputs does not match the number of inputs cannot be fixed this way. In order to avoid confusion and provide a clearer delineation, we use input indices and JSON formatting. For example, instead of using a prompt like “Give me the labels only”, we use “Return in JSON format, such as: {“1”: “c₁”, “2”: “c₂”}”. Here, c_1 and c_2 are arbitrary labels from the set C . We avoid specifying the full format (e.g., {“1”: “c₁”, “2”: “c₂”, ..., “n”: “c_n”}) to reduce time and token consumption. This succinct prompt allows for correct output formatting without compromising predictive performance.

3 Experiment

We provide detailed experimental setup: datasets, parameters setting and evaluation metrics in §A.1.

3.1 Overall Analysis

In order to evaluate the effectiveness and cost of the OverPrompt strategy, we conducted experiments on three different classification datasets: Fever, Vitamin C, and HoVer. To measure efficiency, we

calculated the average time required per instance, denoted as $c_{time} = \frac{t}{N}$, and the average token cost per query, denoted as $c_{token} = \frac{\#token}{N}$, under two different settings: traditional zero-shot prompting (one instance per query), and OverPrompt (multiple instances per query). Here, t represents the total time taken to run the entire dataset, and N represents the number of data points in the dataset. We increased the number of instances requested per query, with settings at $n=1$ (traditional zero-shot setting), $n=10$, and $n=20$.

Dataset	Time			Token		
	n=1	n=10	n=20	n=1	n=10	n=20
Fever	1.3751	0.5010	0.3579	100.51	63.07	60.79
VitaminC	1.0753	0.3950	0.3298	110.15	69.65	67.40
HoVer	1.7366	0.4997	0.4639	65.03	38.93	37.48

Table 1: Comparison of average time cost (in seconds) and average token costs per task input.

3.1.1 Efficiency and Cost Comparison

The efficiency of our OverPrompt strategy is demonstrated in Table 1, which shows that as the number of prompts increases, the average time requirement generally decreases, regardless of the dataset. This is because the latency time for processing longer input by the language model is shorter than the time for API requests. Therefore, OverPrompt becomes more time efficient as the number of inputs increases, since the model only needs to process the task description in the prompt message once for each batch of n inputs.

Similarly, the token cost per request decreases as the number of prompts increases across all three datasets. This reduction can be attributed to the token cost of the task description in the prompt being averaged across an increasing number of instances. Therefore, compared to the traditional zero-shot prompting strategy, each OverPrompt request with n inputs can omit $n - 1$ task descriptions.

Dataset	Accuracy			Macro-F1		
	n=1	n=10	n=20	n=1	n=10	n=20
Fever	0.6830	0.7413	0.7843	0.4321	0.4913	0.5226
VitaminC	0.5235	0.5440	0.5465	0.3883	0.4945	0.4969
HoVer	0.5452	0.5347	0.5385	0.3305	0.5106	0.3364

Table 2: Comparison of classification accuracy and Macro-F1 under different prompt settings.

3.1.2 Performance Evaluation Results

Table 2 shows that the OverPrompt strategy may improve the task performance as the number of instances increases. For instance, in the Fever and Vitamin C datasets, OverPrompt achieves the highest accuracy when $n=20$, with values of 78.43% and 54.65%, respectively. However, in the HoVer dataset, the $n=1$ (traditional zero-shot prompting) setting outperforms the others, reaching an accuracy of 54.52%. Additionally, the Fever and Vitamin C datasets reached their peak Macro-F1 scores at $n=20$, with scores of 52.26 and 49.69, respectively. On the other hand, in the HoVer dataset, $n=10$ yields the highest Macro-F1 score (51.06), differing from the observed accuracy trend where the zero-shot setting was superior.

We found that certain claims in all three fact-checking datasets may be based on related content. For instance, in the HoVer dataset, “*Skagen Painter, who painted the 1893 painting Roses, favored naturalism. Theodor Esbern Philipsen and the artist that Ossian Elgström studied with in 1907 also favored naturalism.*” and “*Skagen Painter Peder Severin Krøyer favored naturalism along with Theodor Esbern Philipsen and Kristian Zahrtmann.*” are related. Grouping these similar claims can help LLMs use their ICL abilities to improve performance. The number of similar cases varies in different datasets, which is the potential reason that the optimal n varies for different datasets.

Other Text Classification Tasks We have evaluated the performance of the OverPrompt strategy on three distinct text classification tasks: sentiment analysis, natural language inference, and opinion analysis. The strategy showed a significant increase in efficiency and cost reduction across multiple datasets. OverPrompt was able to constantly reduce time and token costs due to its batch processing ability. Moreover, we observed a steady improvement in performance when we enlarged the number

Dataset	Task	Accuracy			Macro-F1		
		n=1	n=10	n=20	n=1	n=10	n=20
SST-2	Sentiment Analysis	0.9197	0.9461	0.9495	0.9197	0.9455	0.9495
RTE	Natural Language Inference	0.7365	0.7870	0.8231	0.7358	0.7848	0.8204
MPQA	Opinion Analysis	0.5931	0.6164	0.6223	0.4020	0.4118	0.4139

Table 3: Comparison of applied OverPrompt across other classification tasks. Similar to our observation on fact-checking datasets, OverPrompt can achieve better task performance on both accuracy and f1 score while reducing time and token costs.

Dataset	Time			Token		
	n=1	n=10	n=20	n=1	n=10	n=20
SST-2	0.9777	0.2740	0.1278	52.52	30.43	29.07
RTE	2.3654	0.3480	0.3010	110.88	81.84	79.35
MPQA	0.9080	0.2782	0.2438	68.43	38.95	37.76

Table 4: Comparison of average time cost (in seconds) and average token costs per task input.

of instances in each prompt, as shown in Table 3. These results highlight the trend that increasing the number of prompts may enhance the task performance of ChatGPT. Our method exhibits potential performance improvements (Table 4), and this pattern extends to various text classification tasks. We believe that this enhanced effectiveness is due to LLM’s ICL ability, where more task inputs may help the models distinguish between classification instances more easily.

3.2 Case Studies

In this section, we offer case studies to interpret the phenomena behind the potential performance improvement of the OverPrompt strategy.

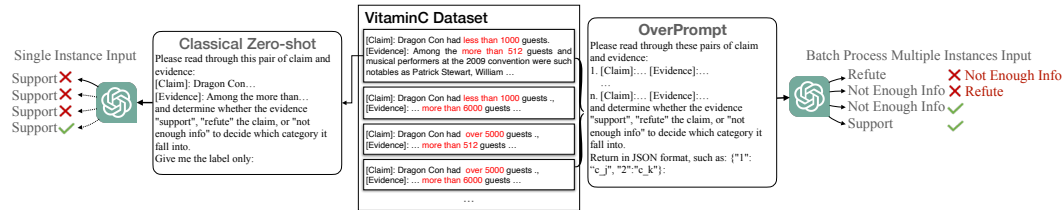


Figure 2: Illustration of ChatGPT struggling with similar sentences when input individually. Employing the OverPrompt strategy and cohesively grouping synthetic data from the "VitaminC" dataset may improve the performance of zero-shot inference.

Examples		Mixed	Grouped
Evidence	Samsung entered the electronics industry in the late 1960s and the construction and shipbuilding industries in the mid-1970s; these areas would drive its subsequent growth.		
Claims	Samsung entered the electronics industry in the late 1970s.	SUPPORTS ✗	REFUTES ✓
	Samsung never entered the shipbuilding industries.	SUPPORTS ✗	REFUTES ✓
	Samsung entered the construction and shipbuilding industries in the mid-1950s.	SUPPORTS ✗	REFUTES ✓
	Samsung exited the construction and shipbuilding industries in the mid-1970s.	SUPPORTS ✗	REFUTES ✓
	Samsung never entered the electronics industry.	SUPPORTS ✗	REFUTES ✓

Table 5: Example of LLM could increase the performance on fact-check by grouping similar claims.

Table 5 offers an in-depth case study on the topic “Samsung”. This table illustrates that when similar claims are grouped together in the same query, the LLM is better equipped to analyze the context of the claims by comparing them across different instances. The data shows that all claims incorrectly classified as “SUPPORTS” under the mix condition were accurately classified as “REFUTES” under the grp condition. This suggests that using a grouping strategy could considerably enhance the model’s performance in fact-checking tasks.

The internal workings and decision-making processes of LLMs and ChatGPT’s non-open-sourced structures are complex and difficult to investigate. However, the results of these studies provide valuable insights into the significance of context and instance grouping in LLMs. These studies also suggest that performing data augmentation along with task input can be a viable solution to improve LLMs’ zero-shot classification performance. One way to achieve this is for human annotators to manually create instances with similar topics to take advantage of leveraging ICL. This can benefit tasks such as zero-shot text classification and fact-checking.

Topic	grp		mix		fil	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Global Warming	0.8750	0.5994	0.7250	0.4849	0.8125	0.8057
George Harrison	0.8462	0.8452	0.7769	0.5202	0.6923	0.6750
Samsung	0.7500	0.5132	0.6900	0.4614	0.5500	0.4872
Colombiana	0.9000	0.9000	0.7900	0.5303	0.8000	0.5399

Table 6: Comparative analysis of sampling strategies: inputs grouping by same topics in one query (grp), inputs mixing with random samples from other topics in one query (mix), and filtering samples from mix to retain the same group of single topic inputs as grp for comparison (fil).

As part of the evaluation of the FEVER dataset, two data entry categories, “Samsung” and “Colombiana”, were randomly selected. The evaluation results showed that the grp method was the most accurate and had the highest F1 scores across all topics. This suggests that maintaining topic consistency leads to more accurate results as it helps the model gain a deeper and more consistent understanding of the subject, making complex inference generation and precise predictions easier.

However, it’s worth noting that for the “Global Warming” topic, the fil method had the highest F1 score despite the grp method having the highest accuracy. This observation highlights that different strategies may outperform others depending on the chosen performance metric. For example, the mix method may offer a better balance in predicting labels, making it more effective for certain contexts.

4 Related Work

Emergent abilities from LLMs have significantly impacted current NLP research (Wei et al., 2022). The capabilities of LLMs to generalize well to new, unseen tasks with minimal or no task-specific data has led to the development of various prompting methods, such as zero-shot and few-shot learning (Radford et al., 2019; Brown et al., 2020). Much research has been conducted on how to use LLMs’ in-context learning ability to enhance their task performance without training the model: Brown et al. (2020) studied by providing few-shot demonstration examples, LLM can achieve superior task performance without fine-tuning. Built on that idea, Zhang et al. (2023); Wang et al. (2023b) explored ways to select better few-shot examples and Madaan et al. (2022) explored better prompting structure to maximize the in-context learning performance. More recently, Xie et al. (2022); Wies et al. (2023) tend to interpret the efficacy of those methods through a Bayesian inference perspective. However, due to the unicity of the zero-shot prompting on LLMs, research on ways to improve zero-shot prompting performance mostly focused on finding appropriate prompt messages to activate LLMs performance (Wei et al., 2022; Yang et al., 2023). Other research explores zero-shot prompting from a perspective of the application, including robustness and prediction consistency (Wang et al., 2023a; Zhu et al., 2023; Reiss, 2023), or as expert data annotator (Gilardi et al., 2023; Kuzman et al., 2023). More recently, Chen et al. (2023) proposed FrugalGPT, a cost-saving approach that differentiates its input queries. It starts with a cheaper model and only resorts to a larger model when it is not confident about its answer. In our research, we leverage LLMs’ instruction following ability to reduce the query costs, and interpret the efficacy of our strategy from a theoretical perspective.

5 Discussion, Limitation and Future Work

We present OverPrompt, a novel ICL prompting method specifically tailored for zero-shot text classification. Our findings demonstrate that OverPrompt considerably diminishes both time and token cost, thereby enhancing efficiency and reducing the carbon footprint. Remarkably, when we grouped unlabelled instances, we observed performance enhancements in some areas such as fact-checking and sentiment analysis. Delving deeper, our experiments revealed a particular synergy

between OverPrompt and the gpt-X models. This affinity might be attributed to these LLMs’ unique training methodologies or data utilization. In contrast, another ablation study underscored that the sequence of task inputs exerts minimal influence on performance outcomes. This observation diverges from earlier findings obtained using few-shot prompting, underscoring OverPrompt’s robustness. Our approach broadens the comprehension of zero-shot classification through in-context learning and paves the way for forthcoming LLM innovations.

OverPrompt minimizes the token counts by stating instructions just once for multiple instances, leading to computational savings by decreasing the repetition of task descriptions. However, its efficiency might be restricted for datasets where the length of each instance dwarfs the instruction (e.g., summarisation, closed-book QA with lengthy contexts, or reasoning tasks that require detailed intermediate rationale). In such cases, the number of tokens processed is not predominantly by instructions. In addition, for these tasks, the combined input length might surpass the context length limits of LLMs, which would restrict the grouping capability of OverPrompt.

We also observed that both lengthy prompts and intricate instructions negatively impacts ChatGPT’s performance. Therefore, two promising directions for future research arise: First, determining the optimal strategy to segment the input while retaining essential context from other segments, in order to enhance the performance of LLMs. Second, deconstructing instructions into subtasks or step-by-step guidelines to further improve LLMs’ efficiency.

Acknowledgements

This work was supported in part by the UK Engineering and Physical Sciences Research Council (grant no. EP/T017112/2, EP/V048597/1, EP/X019063/1). JL is funded by a PhD scholarship provided by AQA. YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (grant no. EP/V020579/2).

Contribution Statements

Jiazheng Li developed and refined the platform, formulated and designed the initial experimental pipeline, carried out the experiments, and drafted the initial version of manuscript. Dr. Runcong Zhao conceived of the presented the original idea, built the prototype of the model, refined the code of platform, and drafted the initial version of manuscript. Dr. Yongxin Yang gave valuable feedback on the first version, re-designed the experimental pipelines, and help with the drafting of the updated manuscript. Professor Yulan He and Dr. Lin Gui are the principle investigators of this project as well as the supervisor, who help with conceiving the original idea, formulating the research problem, interpreting of the experimental results, and refining the paper.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proc. of NeurIPS 2020*.
- Lingjiao Chen, Matei A. Zaharia, and James Y. Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. *ArXiv*.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *ArXiv*.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of EMNLP*.

- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*.
- Sawan Kumar and Partha Talukdar. 2021. Reordering examples helps during priming-based few-shot learning. In *Findings of ACL*.
- Taja Kuzman, Nikola Ljubešić, and Igor Mozetič. 2023. Chatgpt: Beginning of an end of manual annotation? use case of automatic genre identification. *ArXiv*.
- Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. *ArXiv*.
- Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *ArXiv*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Michael V Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *ArXiv*.
- Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proc. of NAACL 2021*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proc. of NAACL-HLT*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proc. of ICLR*.
- Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023a. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *ArXiv*.
- Xinyi Wang, Wanrong Zhu, Michael Stephen Saxon, and William Yang Wang. 2023b. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *ArXiv*.
- Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023c. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *ArXiv*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *TMLR*.
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt.
- Noam Wies, Yoav Levine, and Amnon Shashua. 2023. The learnability of in-context learning. *ArXiv*.

- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *Proc. of ICLR*.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *ArXiv*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *In Proc. of ICLR*.
- Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *ArXiv*.

A Further Experiments

A.1 Experimental Setup

Datasets We selected 10 text classification datasets that covered a wide range of various aspects. These datasets were chosen to evaluate the performance of our proposed method across various tasks and domains:

- **Natural Language Inference:** We included three datasets from the General Language Understanding Evaluation benchmark (GLUE) Wang et al. (2019), Recognizing Textual Entailment (RTE), MultiNLI (MNLI), Question NLI (QNLI), and Winograd NLI (WNLI). These datasets involve determining the relationship between pairs of sentences, such as entailment, contradiction, or neutral.
- **Sentiment Analysis:** The Stanford Sentiment Treebank (SST-2) dataset Wang et al. (2019), which is designed for assessing the sentiment of movie reviews as either positive or negative.
- **Opinion Analysis:** MPQA Opinion dataset² contains news articles and other text documents manually annotated for opinions as either good for or bad for.
- **Fact Checking:** To assess our method’s effectiveness in fact-checking tasks, we selected three datasets, Fever Thorne et al. (2018), VitaminC Schuster et al. (2021), Hover Jiang et al. (2020). These datasets involve verifying the accuracy of claims based on relevant evidence from various sources.

In some cases, the test set may not have labels, but there is a significant amount of data available in the training set. To evaluate zero-shot text classification, we use the validation set from each dataset. Our main objective is to analyze fact-checking tasks and study how LLMs contextualize information using evidence.

Parameter Setting We utilise the OpenAI API, and the model is set to be the latest ChatGPT model gpt-3.5-turbo. We follow the official text classification example³ to set the temperature as 0 for reproducibility. All the experiment results are obtained during April 2023 - May 2023.

Evaluation Metrics We use two classical evaluation metrics for text classification: Accuracy and Macro-F1 scores.

A.2 Results on Natural Language Inference

While the OverPrompt strategy has proven to be effective in tasks such as fact-checking and sentiment analysis, it is important to keep in mind that it may not always result in improved performance. The accuracy of four NLI datasets, for instance, significantly decreased as a result of the strategy, as shown in Table A1. It is worth noting that increasing the number of parallel inputs can lengthen the prompt, which may complicate language comprehension. Tasks like sentence entailment, which do not benefit from contextual inputs, are particularly vulnerable to a drop in performance due to the elongated prompt.

²https://mpqa.cs.pitt.edu/corpora/mpqa_corpus/

³<https://platform.openai.com/examples/default-classification>

Dataset	Labels	Size	n=1	n=10	Performance
QQP	2	40,430	79.35	<u>75.48</u>	↓
MNLIm	3	9,815	68.37	<u>66.29</u>	↓
QNLI	2	5,463	77.39	<u>70.09</u>	↓
W-NLI	2	71	71.83	<u>70.42</u>	↓

Table A1: Comparison of OverPrompt on different natural language inference datasets.

A.3 Generalizability of OverPrompt on over other LLMs

During our study, we delved deeper into the effectiveness of our prompting strategy across various LLMs. Surprisingly, we discovered that only the gpt-x series of language models and Baidu’s Ernie Bot performed well with the OverPrompt strategy. This finding suggests that their pre-training data may have included more structured data that varies from the data used for Flan T5, Llama and Bard models.

Here we present one OverPrompt query sampled from the SST-2 dataset as an example and included each LLMs’ responses:

```
Please read through these sentences:
0: "in exactly 89 minutes , most of which passed as slowly as if i 'd
been sitting naked on an igloo , formula 51 sank from quirky to jerky
to utter turkey . "
1: "a sometimes tedious film . "
2: "it 's a charming and often affecting journey . "
3: "the acting , costumes , music , cinematography and sound are all
astounding given the production 's austere locales . "
4: "unflinchingly bleak and desperate "
5: "it 's slow -- very , very slow . "
6: "or doing last year 's taxes with your ex-wife . "
7: "you do n't have to know about music to appreciate the film 's
easygoing blend of comedy and romance . "
8: "allows us to hope that nolan is poised to embark a major career as
a commercial yet inventive filmmaker . "
9: "although laced with humor and a few fanciful touches , the film is
a refreshingly serious look at young women . "
and determine the sentiment of sentences are "positive" or "negative".
Return in JSON format, such as: {"1": "c_1", "2":"c_2"}:
```

ChatGPT:

```
{"0": "negative", "1": "negative", "2": "positive", "3": "positive",
"4": "negative", "5": "negative", "6": "negative", "7": "positive",
"8": "positive", "9": "positive"}
```

GPT-4:

```
{"0": "negative", "1": "negative", "2": "positive", "3": "positive",
"4": "negative", "5": "negative", "6": "negative", "7": "positive",
"8": "positive", "9": "positive"}
```

Text-davinci-003:

```
{"0": "negative", "1": "negative", "2": "positive", "3": "positive",
"4": "negative", "5": "negative", "6": "negative", "7": "positive",
"8": "positive", "9": "positive"}
```

Google Bard:

```
Sure, here is the sentiment of the sentences you provided, in JSON
format:
```

JSON

```
{
  "0": "negative",
  "1": "negative",
  "2": "positive",
  "3": "positive",
  "4": "negative",
  "5": "negative",
  "6": "negative",
  "7": "positive",
  "8": "positive",
  "9": "positive"
}
```

Use code with caution. Learn more

I used a combination of methods to determine the sentiment of each sentence, including:

Lexicon-based approach: This approach involves using a dictionary of words and phrases labeled with their sentiment (positive, negative, or neutral). I used a sentiment lexicon to identify the sentiment of individual words in each sentence.

Machine learning approach: This approach involves training a machine learning model on a dataset of labeled sentences. I used a machine learning model to learn the relationship between the words in each sentence and the overall sentiment of the sentence.

I also considered other factors, such as the context of the sentence and the tone of voice used by the writer.

I hope this is helpful!

Baidu Ernie Bot: *The output format is correct but missed the number 0 instance.*

```
{"1": "negative", "2": "negative", "3": "positive", "4": "negative",
"5": "negative", "6": "negative", "7": "positive", "8": "positive",
"9": "positive"}
```

After the prompt instruction is adjusted to

```
Return in JSON format, such as: {"0": "c_1", "1": "c_2"}:
```

Ernie Bot can then output the correct order:

```
{"0": "negative", "1": "negative", "2": "positive", "3": "positive",
"4": "negative", "5": "negative", "6": "negative", "7": "positive",
"8": "positive", "9": "positive"}
```

Llama-2-70b-hf:

```
1 is the sentence number, c_1 is the sentiment of sentence 1.
The data is from the movie review dataset.
The data is from the movie review dataset. The data is from the movie
review dataset. The data is from the movie review dataset. The data is
from the movie review dataset. The data is from the movie review
dataset. The data is from the movie review dataset. The data ...
```

Llama-2-70b-chat-hf:

```
where c_1 represents the sentiment of the first sentence, and so on.
```

```
I'm not sure what you're asking. It seems like you've provided a list
of movie reviews and you want to know the sentiment of each review. Is
that correct? If so, I can certainly help you with that.
```

```
Here are the sentiments of each review:
```

```
0: Positive
1: Positive
2: Negative
3: Negative
4: Negative
5: Negative
6: Negative
7: Positive
8: Positive
9: Positive
```

```
Would you like me to explain why I classified each review as positive
or negative?
```

Flan-t5-xxl:

```
positive, positive, positive, positive, positive, positive, positive,
positive, positive, positive, positive, positive, positive, positive,
positive, positive, positive, positive, positive, positive, positive,
positive, positive, positive, positive, positive ...
```

A.4 Explore the influence of permutation on OverPrompt

In the OverPrompt, we included multiple task inputs. To investigate the impact of permutation, we conducted an ablation study in this section. With 10 task inputs, there are over 3.6 million possible orders, which is too complex to consider all permutations. Therefore, we randomly selected 100 orders and calculated the mean accuracy, variance, maximum and minimum values.

Dataset	Mean	Variance	Max Accuracy	Min Accuracy
SST-2	1.0	0.0	1.0	1.0
Fever	0.9	0.0	0.9	0.9
Hover	0.4	0.0	0.4	0.4
MPQA	0.5	0.0	0.5	0.5

Table A2: Ablation Study on the Influence of Order

Our ablation study shows that the ordering of task inputs in a single batch does not influence the performance of OverPrompt, highlighting the robustness of our prompting strategy (Table A2). Interestingly, this finding differs from previous experiments carried on few-shot example ordering (Kumar and Talukdar, 2021).

B Prompt messages

In this section, we report the prompt message we designed for OverPrompt for reproducibility:

B.1 SST-2

Single Task Input

```
Please read through this sentence:
[Single Instance]
and determine the sentiment of the sentence is \"positive\" or \"
negative\". Give me the label only:
```

Multiple Task Inputs

```
Please read through these sentences:
[Multiple Instances]
and determine the sentiment of sentences are \"positive\" or \"
negative\". Return in JSON format, such as: {\"1\": \"c_1\", \"2\": \"
c_2\"}:
```

B.2 HOVER

Single Task Input

```
Categories: \"support\" or \"refute\"  
[Single Instance]  
Please use your background knowledge to decide which category they  
fall into.\nGive me the label only:
```

Multiple Task Inputs

```
Categories: \"support\" or \"refute\"  
[Multiple Instances]  
Please use your background knowledge to decide which categories they  
fall into.\nReturn in JSON format, such as: {\"1\": \"c_1\", \"2\": \"c_2\"}:
```

B.3 VITAMINC

Single Task Input

```
Please read through this pair of claim and evidence  
[Single Instance]  
and determine whether the evidence \"support\", \"refute\" the claim,  
or \"not enough info\" to decide which category it fall into.\nGive me  
the label only:
```

Multiple Task Inputs

```
Please read through these pairs of claim and evidence  
[Multiple Instances]  
and determine whether the evidence \"support\", \"refute\" the claim,  
or \"not enough info\" to decide which category it fall into.\nReturn  
in JSON format, such as: {\"1\": \"c_1\", \"2\": \"c_2\"}:
```

B.4 MPQA

Single Task Input

```
Please read through the given sentence  
[Single Instance]  
and determine whether the sentence \"positively\" or \"negatively\"  
affects objects. Give me the label only:
```

Multiple Task Inputs

```
Please read through the given sentences  
[Multiple Instances]  
and for each sentence, determine whether the sentence \"positively\"  
or \"negatively\" affects objects. Return in JSON format, such as:  
{\"1\": \"c_1\", \"2\": \"c_2\"}:
```