

How good are you? An empirical classification performance comparison of Large Language Models with traditional Open Set Recognition classifiers

Alexander Grote^{1,*†}, Anuja Hariharan^{2,†}, Michael Knierim^{2,†} and Christof Weinhardt^{2,†}

¹FZI Research Center for Information Technology, Haid-und-Neu-Str. 10–14, 76131 Karlsruhe, Germany

²Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany

Abstract

The release of ChatGPT has led to an unprecedented surge in the popularity of generative AI-based Large Language Models (LLMs) among practitioners. These models have gained traction in business processes due to their ability to receive instructions in natural language. However, they suffer from hallucinations, which are generated texts that are factually incorrect. Hallucinations also arise in text classification tasks, such as customer support ticket classification or intent classification for chatbots. In such scenarios, the user prompts the model to classify an incoming text into predefined categories. Furthermore, in real-world scenarios, it is common to encounter texts that do not fit into the predefined categories. It is unclear if current state-of-the-art LLM can handle such scenarios and how they compare to existing classifiers focusing on these situations. In this paper, we propose a way to evaluate the classification performance of LLMs in an Open Set Recognition (OSR) scenario, where unseen classes can occur at inference time. The simulation consists of an empirical comparison between GPT4 and Gemini Pro, two state-of-the-art language models, a fine-tuned version of GPT3.5 and established OSR classifiers. The results would provide insights into how reliable large language models are for classification purposes and if they can replace existing OSR classifiers that typically require a decent amount of labelled data.

Keywords

Large Language Models, Open Set Recognition, Classification

1. Introduction

Since the release of ChatGPT in November 2022 [1], the adoption of Large Language Models (LLMs) in businesses has experienced significant growth [2]. Especially the ability to use natural language to interact with these models has allowed practitioners with little programming knowledge to harness the power of such systems in their daily operations. However, utilising LLMs comes at the risk of factually incorrect generated texts, also known as hallucinations [3]. Often, these hallucinations are undesired and, for example in the intent classification used in chatbot interactions, they might negatively impact customer service quality and potentially


16th ZEUS Workshop, ZEUS 2024, Ulm, Germany, February 29th - March 1st, 2024, Germany

*Corresponding author.

†These authors contributed equally.

✉ grote@fzi.de (A. Grote); anuja.hariharan@kit.edu (A. Hariharan); michael.knierim@kit.edu (M. Knierim); weinhardt@kit.edu (C. Weinhardt)

ORCID 0009-0005-9743-6648 (A. Hariharan); 0000-0001-7148-5138 (M. Knierim)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

S. Böhm and D. Lübke (Eds.): 16th ZEUS Workshop, ZEUS 2024, Ulm, Germany, 29 February–1 March 2024, published at <http://ceur-ws.org>

harm the company’s reputation. This highlights the need of a robust system that is not only able to classify the customer intent correctly, but also detects out-of-distribution questions and replies accordingly. The functionality of a system that rejects out-of-distribution data points and classifies known patterns into existing categories has been widely studied under the term Open Set Recognition (OSR) [4]. In particular, deep learning based OSR classifiers, such as the OpenMax [5] or the DOC [6] algorithm, have shown an increased performance on OSR classification tasks [7]. Similarly, the zero- [8] and few-shot [9] abilities of LLMs have also been leveraged to solve these tasks. Due to the fast-paced advancements in the realm of LLMs, it is unclear from a practitioner’s point of view how well state-of-the-art LLMs with zero- and few-shot strategies compare to established solutions from OSR and how reliable they are in a production setting. This ultimately leads to the question of which approach to choose and how they compare against each other. In this work, we plan to provide insights into the classification accuracy and the ability to reject unknown instances by conducting an empirical analysis between these two research areas. We thereby give guidance for practitioners and an updated benchmark for the current state-of-the-art LLM classification performance.

2. Related Work

Generative Pre-trained Transformer (GPT) models represent a paradigm shift in Natural Language Processing (NLP) [10]. While these LLMs are typically pretrained in a self-supervised, task-independent manner, they are known to be very good at NLP tasks, even without fine-tuning [11]. To use these models for classification tasks one can either fine-tune the model or use zero- and few-shot techniques for in-context-learning. Fine-tuning involves adjusting the weights of a pre-trained model for a particular task and, given a large dataset, supersedes the classification performance of zero- and few-shot strategies [12]. In contrast, zero- and few-shot learning methods utilise the capability of Large Language Models (LLMs) to categorise new data points effectively, even when they have encountered none or only a minimal number of examples from a specific class. Typically, zero- and few-shot strategies are combined with prompting strategies, such as ”Chain of Thought” [13] and ”Clue And Reasoning Prompting” [14], to further enhance the classification performance. Despite these strategies, Kocoń et al. [15] and Caruccio et al. [16] have demonstrated that the zero- and few-shot capabilities are worse than supervised machine learning models for classification tasks. In their analysis, however, they assumed a closed set scenario, which is an unrealistic assumption.

A more realistic scenario than traditional classification is Open Set Recognition [4]. It allows for unknown classes during inference, and the classifier has an additional option to reject data points as unknown. If the incoming data point is not rejected as unknown, the classifier classifies the data point into a known class. Among the first OSR models were adapted Support Vector Machines [17, 18]. With the rise of neural networks, Bendale and Boulton [5] reformulated the final softmax layer to also estimate the probability of a data point being out-of-distribution. Similarly, Shu et al. [6] use a one-versus-rest classification layer to reduce the misclassifications in the open space, while Oza and Patel [19] utilise an autoencoder and its reconstruction loss to determine if a data point is novel.

3. Proposed Approach

To compare the performance of LLMs versus Open Set Recognition classifiers, we plan to set up an empirical evaluation as illustrated in Figure 1.

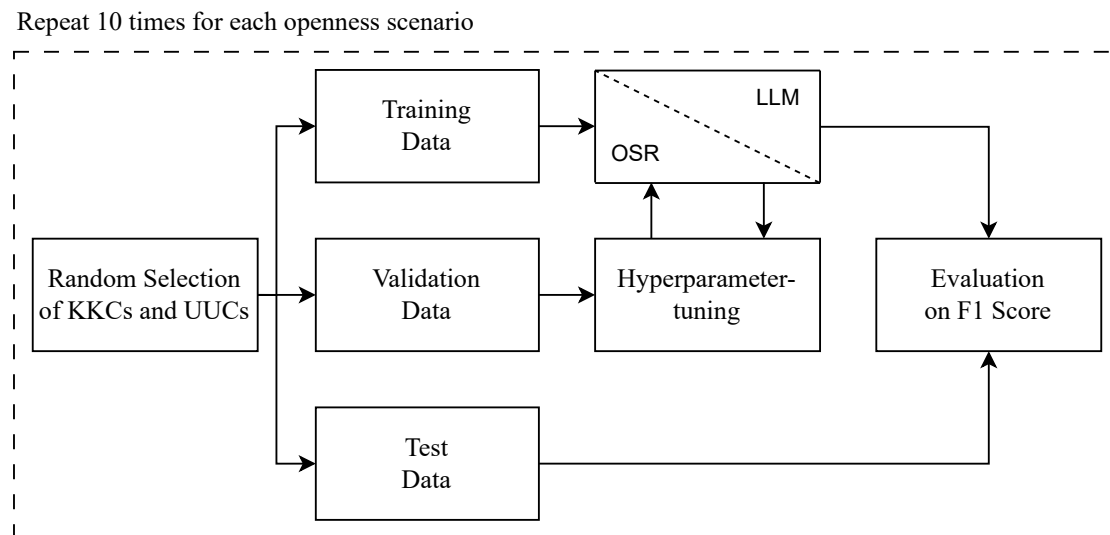


Figure 1: Machine learning workflow to compare conversational LLMs with OSR classifiers.

For our experiments, we will use four different text classification datasets. These datasets include the 20 Newsgroups dataset [20], the Yahoo! Answers dataset [21], the CLINC150 dataset [22] and the BANKING77 dataset [23]. While the first dataset consists of news articles and the second questions-answer pairs of certain categories, the last two represent intent classification tasks. All datasets have at least ten different classes or categories, based on which we will simulate an open set scenario. We follow the data splitting procedure of Geng et al. [7] to select the known and unknown classes for the open set simulation and repeat each simulation ten times to derive statistically meaningful results. Furthermore, we plan to exclude 0, 10, 20 and 30 % of all available classes from training and evaluate the classification for each scenario on the f1 score. The f1 score is a commonly used metric in classification problems, measuring the harmonic mean between precision and recall. However, in OSR scenarios, the unknown classes are typically not considered as an additional class when calculating the f1 score [7]. That is why we additionally distinguish between the f1 score classification performance on the known and unknown classes, providing further insights into the applicability of LLMs for open scenarios. In terms of conversational LLMs, we plan to use two state-of-the-art models, GPT4 [24] from OpenAI and Gemini Pro [25] from Google, with zero-shot and few-shot prompt configurations. When using a zero-shot configuration, we provide the LLM with only the category name and description, while in a few-shot setting, we also include examples of each category. Currently, it is not possible to create a custom, fine-tuned model from both of these two models. Instead, we will use OpenAI’s GPT3.5 model and fine-tune it with resources from OpenAI to also investigate the improvements made through fine-tuning. We then compare the results to the classification

performance of the OpenMax [5] and DOC [6] classifiers. To speed up the training process of both OSR classifiers, we first transform the incoming texts into meaningful embeddings using the most advanced text embedding provided by OpenAI [26] and then train a shallow neural network on the retrieved embeddings. The shallow neural network integrates either the OpenMax or the DOC architecture.

4. Conclusion

Generative AI models for text generation, like ChatGPT, have proven useful in various tasks. In particular, they can classify an incoming text into predefined categories. In this paper, we propose a study design that compares the classification performance of state-of-the-art LLMs with existing classifiers for Open Set Recognition. The results of this study provide insights into the reliability of conversational LLMs and whether they are a viable alternative to traditional classification systems.

References

- [1] OpenAI, Introducing ChatGPT, 2022. URL: <https://openai.com/blog/chatgpt>.
- [2] W. Hariri, Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing (2023). URL: <https://arxiv.org/abs/2304.02017>. doi:10.48550/ARXIV.2304.02017, publisher: arXiv Version Number: 6.
- [3] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, J.-R. Wen, Halueval: A large-scale hallucination evaluation benchmark for large language models, in: Proceedings of the 2023 conference on empirical methods in natural language processing, 2023, pp. 6449–6464.
- [4] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, T. E. Boult, Toward Open Set Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2013) 1757–1772. doi:10.1109/TPAMI.2012.256, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [5] A. Bendale, T. E. Boult, Towards Open Set Deep Networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 1563–1572. URL: <http://ieeexplore.ieee.org/document/7780542/>. doi:10.1109/CVPR.2016.173.
- [6] L. Shu, H. Xu, B. Liu, DOC: Deep Open Classification of Text Documents, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2911–2916. URL: <https://aclanthology.org/D17-1314>. doi:10.18653/v1/D17-1314.
- [7] C. Geng, S.-J. Huang, S. Chen, Recent Advances in Open Set Recognition: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2021) 3614–3631. doi:10.1109/TPAMI.2020.2981604, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [8] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le,

- Finetuned language models are zero-shot learners, in: International conference on learning representations, 2022. URL: <https://openreview.net/forum?id=gEZrGCozdqR>.
- [9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, others, Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [10] T.-X. Sun, X.-Y. Liu, X.-P. Qiu, X.-J. Huang, Paradigm Shift in Natural Language Processing, *Machine Intelligence Research* 19 (2022) 169–183. URL: <https://link.springer.com/10.1007/s11633-022-1331-6>. doi:10.1007/s11633-022-1331-6.
- [11] K. S. Kalyan, A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4, *SSRN Electronic Journal* (2023). URL: <https://www.ssrn.com/abstract=4593895>. doi:10.2139/ssrn.4593895.
- [12] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, C. A. Raffel, Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, *Advances in Neural Information Processing Systems* 35 (2022) 1950–1965.
- [13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, others, Chain-of-thought prompting elicits reasoning in large language models, *Advances in Neural Information Processing Systems* 35 (2022) 24824–24837.
- [14] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, G. Wang, Text Classification via Large Language Models (2023). URL: <https://arxiv.org/abs/2305.08377>. doi:10.48550/ARXIV.2305.08377, publisher: arXiv Version Number: 3.
- [15] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, A. Kocoń, B. Koptyra, W. Mieleśczenko-Kowszewicz, P. Miłkowski, M. Oleksy, M. Piasecki, L. Radlinski, K. Wojtasik, S. Woźniak, P. Kazienko, ChatGPT: Jack of all trades, master of none, *Information Fusion* 99 (2023) 101861. URL: <https://linkinghub.elsevier.com/retrieve/pii/S156625352300177X>. doi:10.1016/j.inffus.2023.101861.
- [16] L. Caruccio, S. Cirillo, G. Polese, G. Solimando, S. Sundaramurthy, G. Tortora, Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot, *Expert Systems with Applications* 235 (2024) 121186. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0957417423016883>. doi:10.1016/j.eswa.2023.121186.
- [17] W. J. Scheirer, L. P. Jain, T. E. Boult, Probability Models for Open Set Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (2014) 2317–2324. doi:10.1109/TPAMI.2014.2321392, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [18] L. P. Jain, W. J. Scheirer, T. E. Boult, Multi-class Open Set Recognition Using Probability of Inclusion, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2014, pp. 393–409. doi:https://doi.org/10.1007/978-3-319-10578-9_26.
- [19] P. Oza, V. M. Patel, C2ae: Class conditioned auto-encoder for open-set recognition, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2307–2316.
- [20] D. Lewis, Reuters-21578 text categorization collection, 1997. Text.howpublished: UCI Machine Learning Repository.
- [21] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification,

- Advances in neural information processing systems 28 (2015).
- [22] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, J. Mars, An evaluation dataset for intent classification and out-of-scope prediction, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1311–1316. URL: <https://aclanthology.org/D19-1131>. doi:10.18653/v1/D19-1131.
 - [23] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, I. Vulić, Efficient intent detection with dual sentence encoders, in: T.-H. Wen, A. Celikyilmaz, Z. Yu, A. Papangelis, M. Eric, A. Kumar, I. Casanueva, R. Shah (Eds.), Proceedings of the 2nd workshop on natural language processing for conversational AI, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://aclanthology.org/2020.nlp4convai-1.5>. doi:10.18653/v1/2020.nlp4convai-1.5.
 - [24] OpenAI, GPT-4 Technical Report (2023). URL: <https://arxiv.org/abs/2303.08774>. doi:10.48550/ARXIV.2303.08774, publisher: arXiv tex.version: 4.
 - [25] Gemini Team, Gemini: A Family of Highly Capable Multimodal Models (2023). URL: <https://arxiv.org/abs/2312.11805>. doi:10.48550/ARXIV.2312.11805, publisher: arXiv tex.version: 1.
 - [26] OpenAI, New and improved embedding model, 2022. URL: <https://openai.com/blog/new-and-improved-embedding-model>.