

Using Small Language Models For Local Email Categorization Into Variable Number Of User-Defined Labels

1st Jakub Matłacz (Author)
*Student of Applied Computer Science,
Data Engineering Speciality
Faculty Of Electrical Engineering
Warsaw University Of Technology
Warsaw, Poland
jakub.matlacz.stud@pw.edu.pl*

2nd Krzysztof Hryniów (Supervisor)
*Institute of Control and Industrial Electronics
Faculty Of Electrical Engineering
Warsaw University Of Technology
Warsaw, Poland
krzysztof.hryniow@pw.edu.pl*

Abstract—A new method for local categorization of e-mail messages using small language models was presented. The method consists of three main stages: analysis, selection, and matching (ASM). The goal was to allow the user to define their own labels of variable quantity and quality, which is the distinguishing feature of this work. A collection consisting of one hundred messages with unique classes was prepared. A combination of a language model and a prompt engineering technique was chosen to give the best result in the task of guessing the true class from among a random set of potential classes containing the correct class. Five different models were tested, sized to allow wide adaptation of the solution on consumer computers. The best approach achieved a score of 97% accuracy. The final product met all assumptions and featured high accuracy, speed, and small context size in the model's query.

Index Terms—Analyze-Select-Match (ASM), email categorization, small language models, text categorization/classification, Gestalt Pattern Matching, prompting techniques, natural language processing, artificial intelligence

I. INTRODUCTION

The project aimed to create a system that allows the private classification of messages on the user's local machine. The user could define their categories, and the model would match the message to one of them. Best of all, it does not require the model to be fine-tuned. A dataset consisting of a hundred messages, a true class, and a list of random potential classes of variable length containing the correct class was created.

To accommodate most users' graphics cards in VRAM, five distinct models (Table I) were tested [1]. Eight additional approaches were also investigated; five distinct prompts were used, including proprietary formulations, and various prompt engineering strategies. The greatest accuracy of 97% was attained. Performance, however, which was determined by comparing the accuracy to the average time required to categorize a single message, was just as important.

In the end, an effective technique that most hardware users could utilize locally was found. With no retraining, the model performs as an agent, greatly benefiting seamless transitions between various tasks. Consequently, a novel approach was

presented: Analyze-Select-Match, or ASM for short (Figure 1).

II. METHODS

Since no publicly available datasets nearly matched the project's requirements, a custom dataset had to be created. A dataset was created with messages, a list of categories, of which only one was real, and a valid category. The list's number of categories varied and was chosen randomly from those already included in the larger dataset.

The email landscape consisted of around ten main categories, with ten subcategories assigned to each main category. After that, one example message was written for every subcategory, creating hundred messages with distinct categories.

Using information available on the Steam platform, the distribution of VRAM among its user population was determined (Figure 2). Then, only language models with appropriately small sizes were considered, allowing for a quick and localized deployment, which was in line with the project's primary requirements.

Among a set of suitably scaled models, five that the author personally selected were among the well-known and well-accepted models in the community. Testing was done on each model using the same set of prompt engineering approaches and prompts, considering accuracy and temporal metrics.

The models' default settings, which the developers of each model carefully calibrated, were followed (Table II) with sporadic modifications restricted to reaction length, temperature, and context size. LM Studio [2], Python, and the diffllib library [3] were among the tools that were wisely used during the project. The classifying procedure usually happens in phases. First, the model carefully examined the message, producing additional ideas and related information. It then replied with a single word and chose a category from the ones the user had supplied.

Next, to choose the class that most closely resembled one of the options, the Gestalt Pattern Matching algorithm [4] was

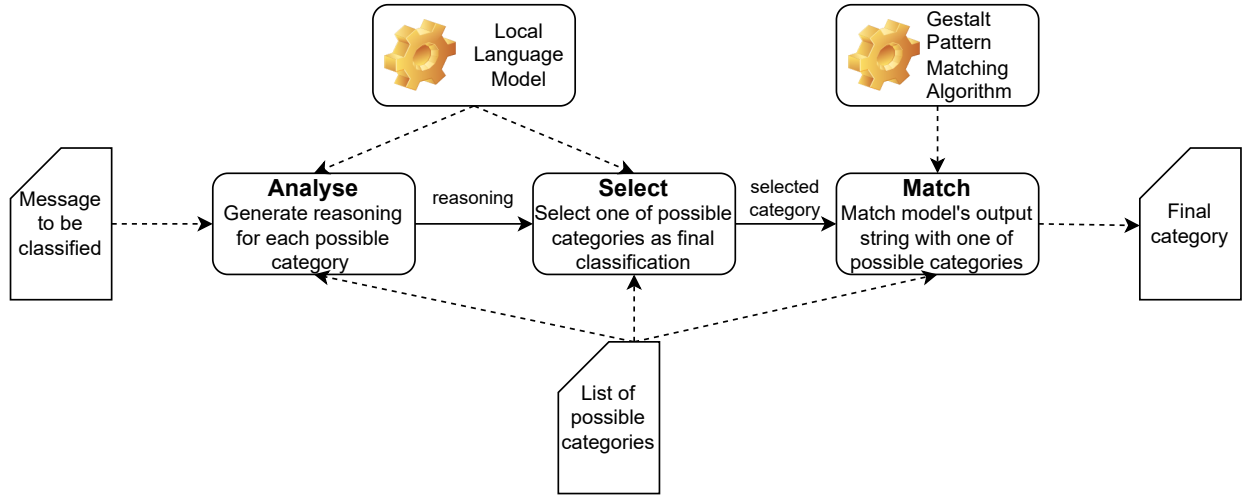


Fig. 1. Graph of ASM method's data processing.

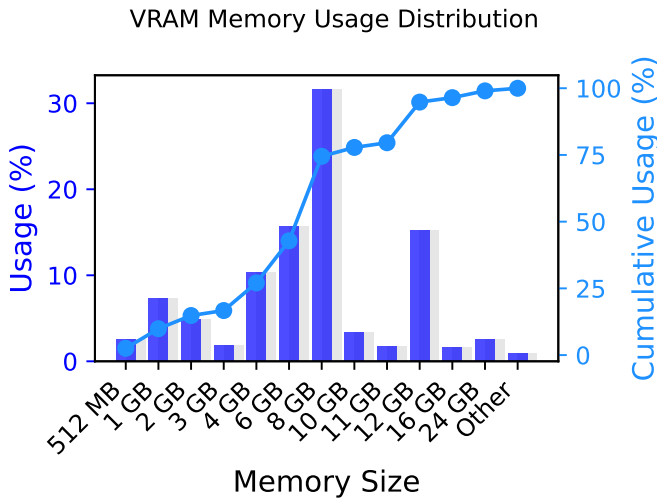


Fig. 2. Distribution of VRAM amount on Steam users' computers [1].

Model Name	Quantization	Context Length	Parameters
ehartford_dolphin-2.0-mistral-7b [5]	Q6_K	32768	7B
teknium_openhermes-2.5-mistral-7b [6]	Q6_K	32768	7B
Phi2 [7]	Q6_K	2048	3B
mistralai_mistral-7b-instruct-v0.2 [8]	Q6_K	32768	7B
open-orca_mistral-7b-openorca [9]	Q6_K	32768	7B

TABLE I
TECHNICAL DETAILS OF USED LANGUAGE MODELS.

used. This method significantly improved the findings because models frequently reply succinctly or slightly alter the original class names. The process known as Analyze-Select-Match, or ASM for short, was used (Picture 1).

Inference Parameters	Value
Temperature (temp)	0.8
Tokens to Generate (n_predict)	-1
Top K Sampling (top_k)	40
Repeat Penalty (repeat_penalty)	1.1
Min P Sampling (min_p)	0.05
Top P Sampling (top_p)	0.95

TABLE II
DEFAULT INFERENCE PARAMETERS THAT WERE USED IN EXPERIMENTS WITH ALL LANGUAGE MODELS.

III. RESULTS

Figure no 3 shows the average metrics of the various querying techniques in the different models. The COT_1 technique had the highest average accuracy among the models tested. The highest average performance and speed, on the other hand, had the IO_1 technique.

In figure no 4, you can see a heat map showing the experimental accuracies in different approaches, that is, combinations of model and query engineering techniques. The highest score (0.97) was achieved by the combination of COT_1 and OPENHERMES MISTRAL 2.5 7B. It can be seen that the PHI 2 3B model [10] performed the worst of all, although it had relatively good accuracy in the COT_2 approach. The model with the highest stability when testing various query engineering techniques was MISTRAL INSTRUCT 0.2 7B. The techniques COT_2, IO_1, IO_1 W/ CS-5 characterized the worst accuracy.

Figure no 5 shows a heat map showing the experimental average query processing times of single messages in different approaches. The fastest query engineering techniques were IO_1, IO_1 W/ CS-5, while the slowest was PHI 2 3B combined with COT_3, 3-SHOT [11] COT_3 techniques (implementation of CARP technique [12]). Input-Output (IO) techniques [13], while efficient, exhibited lower accuracy. The Chain-of-Thought (COT) [?] approach might be less efficient but exhibits high accuracy. The 3-SHOT COT_1 technique had the greatest temporal stability in model variability. The COT_1

technique was characterized by time stability. The combination of COT_1 and OPENHERMES MISTRAL 2.5 7B achieved an average time of 8.40 seconds. The combination of 3-SHOT COT_3 and DOLPHIN MISTRAL 2.0 7B achieved a good time of 5.03 seconds and a good accuracy score 0.91.

Figure no 6 shows a heat map showing the experimental efficiencies, understood as accuracy divided by time, in different approaches. The IO_1 and IO_1 W/ CS-5 techniques characterize the highest efficiency. PHI 2 3B generally characterized the worst efficiency. The combination of COT_1 and OPENHERMES MISTRAL 2.5 7B had a relatively good efficiency of 0.12. Combining 3-SHOT COT_3 and DOLPHIN MISTRAL 2.0 7B achieved good efficiency (0.18) with reasonable accuracy (0.91).

Overall, the best approach combines COT_1 and OPENHERMES MISTRAL 2.5 7B, which achieved an accuracy score of 0.97, an efficiency of 0.12, and a processing time of 8.40. However, in the end, the choice should depend on the specific requirements because, as you can see, there is a trade-off between accuracy and processing speed. The Mistral model and its derivatives presented strong results [14].

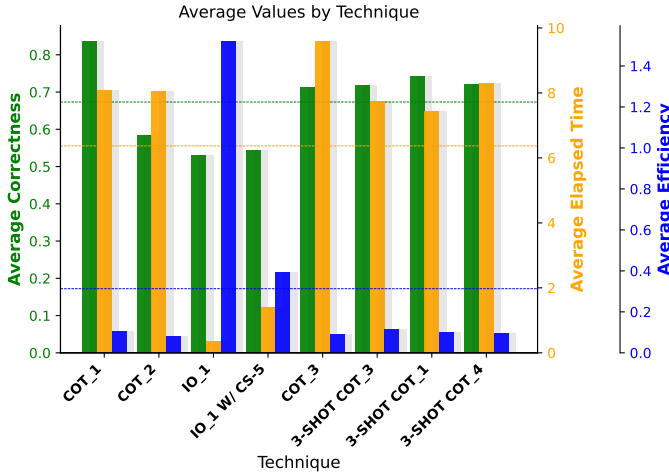


Fig. 3. Average metrics of different prompting techniques in different models.

IV. CONCLUSIONS

The success of the solution we created lies in how well it matches the predefined expectations for categorizing local e-mail messages. It consistently achieves high accuracy and efficiently organizes a wide variety of messages. Processing time is fast, and performance is high. The efficiency and resilience of the solution are high. In addition to classification, this model can switch between tasks smoothly, as it was not fine-tuned to any dataset. The solution is easy to use in a production environment. Indeed, the solution is highly available, can be relied upon, and uses minimal resources. The options allow users to choose their own categories, which can be very specific to them. The model presents excellent accuracy and performance. It can be relied upon and is focused on the user's needs.

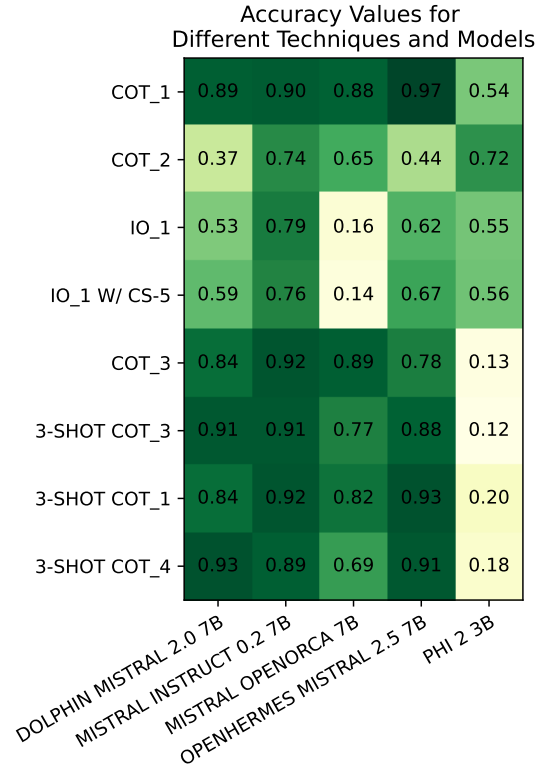


Fig. 4. Experimental accuracies in different approaches.

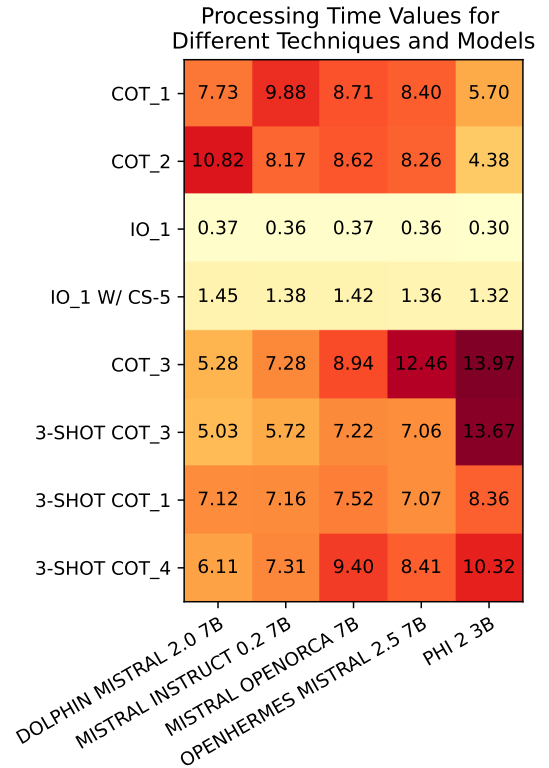


Fig. 5. Experimental average processing times of single message in different approaches.

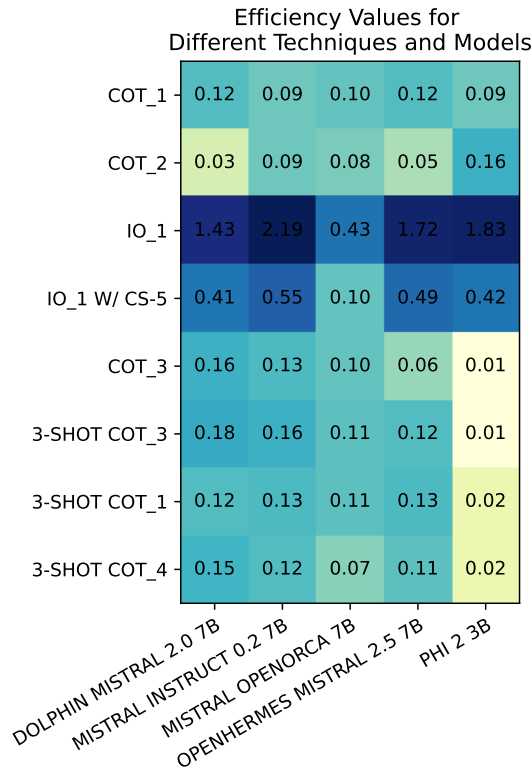


Fig. 6. Experimental efficiencies (accuracy/time) in different approaches.

V. FUTURE WORK

A future solution to improving this model could focus on cybersecurity so that a person with bad intentions cannot present a spam message and make us believe it is a message from one of our categories. In addition, the focus could be on improving the user experience based on data collected from the user's interactions with the system. We could increase the dataset size on which combinations of models and language model query engineering techniques were tested. We could improve energy efficiency to protect the environment. In addition, more advanced classical algorithms in NLP with high energy efficiency could be applied. All in all, the future of this project is focused on its continuous improvement, and it is about the quality interpreted by the user and its environmental and safety performance.

REFERENCES

- [1] Valve Corporation. (2024) Steam Hardware & Software Survey. Accessed on January 11, 2024. [Online]. Available: <https://store.steampowered.com/hwsurvey/Steam-Hardware-Software-Survey-Welcome-to-Steam>
- [2] LM Studio, "LM Studio GitHub Repository," <https://github.com/lmstudio-ai>, 2023, accessed on January 11, 2024.
- [3] Python Software Foundation, *difflib – Helpers for computing deltas*, 2023, <https://docs.python.org/3/library/difflib.html>.
- [4] Wikipedia contributors. (2023) Gestalt pattern matching. [Online]. Available: https://en.wikipedia.org/wiki/Gestalt_pattern_matching
- [5] TheBloke. (2023) TheBloke/dolphin-2.0-mistral-7B-GGUF. Accessed on January 11, 2024. [Online]. Available: <https://huggingface.co/TheBloke/dolphin-2.0-mistral-7B-GGUF>

- [6] ——. (2023) TheBloke/OpenHermes-2.5-Mistral-7B-GGUF. Accessed on January 11, 2024. [Online]. Available: <https://huggingface.co/TheBloke/OpenHermes-2.5-Mistral-7B-GGUF>
- [7] ——. (2023) TheBloke/phi-2-GGUF. Accessed on January 11, 2024. [Online]. Available: <https://huggingface.co/TheBloke/phi-2-GGUF>
- [8] ——. (2023) TheBloke/Mistral-7B-Instruct-v0.2-GGUF. Accessed on January 11, 2024. [Online]. Available: <https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF>
- [9] ——. (2023) TheBloke/Mistral-7B-OpenOrca-GGUF. Accessed on January 11, 2024. [Online]. Available: <https://huggingface.co/TheBloke/Mistral-7B-OpenOrca-GGUF>
- [10] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi *et al.*, "Textbooks are all you need," *arXiv preprint arXiv:2306.11644*, 2023.
- [11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [12] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang, "Text classification via large language models," *arXiv preprint arXiv:2305.08377*, 2023.
- [13] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," 2022.
- [14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. I. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

VI. PROMPTS AND OTHER DETAILS

Project's Repository

Example Prompts

Dataset

Experiment's Configuration File

VII. PROMPTS

All prompts can be accessed inside the project's GitHub repository. Below are visible some of the most important examples for immediate access.

Example IO_1 prompting response of MISTRAL INSTRUCT 0.2 7B model:

"Q: What's the immediate one-word categorization for the message: 'Quality sleep is key to overall well-being. Establish a bedtime routine and prioritize 7-9 hours of rest each night for optimal health.'? Stick to the provided categories: '1. sleep 2. cultural_appreciation 3. cultural_events 4. relaxation_and_stress_management 5. friendships 6. adaptability_and_resilience 7. hobbies 8. savings 9. ethical_decision_making'."

"A: relaxation."

Example COT_4 prompting response of DOLPHIN MISTRAL 2.0 7B model:

" Develop an effective TOPIC CLASSIFICATION SYSTEM for the given INPUT by initially presenting relevant CLUES such as KEYWORDS, PHRASES, CONTEXTUAL INFORMATION, SEMANTIC RELATIONS, TONE, and REFERENCES. Subsequently, establish a diagnostic REASONING PROCESS based on CLUES to support the determination of the TOPIC. Finally, discern the overarching TOPIC of the INPUT within predefined CATEGORIES by integrating the identified CLUES, the established REASONING PROCESS, and the INPUT itself. YOUR TASK: INPUT: - MESSAGE: Explore new recipes and ensure your diet is well-balanced with a mix of fruits, vegetables, and proteins. Your nutrition is the foundation of good health. - CATEGORIES: 1. professional_development 2. financial_education 3. nutrition 4. conflict_resolution 5. formal_education 6. empathy_and_understanding Now think about the CLUES, REASONING, and OUTPUT for this message. "

" CLUES: - KEYWORDS: explore new recipes, well-balanced diet, fruits, vegetables, proteins, nutrition, good health. - PHRASES: nutrition as the foundation of good health. - CONTEXTUAL INFORMATION: nutrition as the foundation of good health. - SEMANTIC RELATIONS: exploring new recipes and maintaining a balanced diet. - TONE: positive. - REFERENCES: nutrition, good health, balanced diet, recipes. REASONING: The KEYWORDS and PHRASES, such as explore new recipes, a well-balanced diet, and nutrition as the foundation of good health, suggest a focus on maintaining a balanced diet and nutrition. The mention of a balanced diet consisting of fruits, vegetables, and proteins further supports the idea of nutrition as a key aspect of good health. OUTPUT: nutrition OUTPUT: The overarching TOPIC of the INPUT is nutrition."

"Q: Thank you for your reasoning! We need an answer in very strict format so that our automatic systems will be able to process it. It is very important. Please tell what is the final verdict of your expert categorization. Answer in one word. Only use word from the potential TOPICs. Any deviation from these rules will result in system crash. A: The final verdict in single word is: "

" Nutrition"

Example COT_3 prompting response of MISTRAL INSTRUCT 0.2 7B model:

"This is an overall topic classifier for MESSAGE. First, present CLUES (i.e., keywords, phrases, contextual information, semantic relations, semantic meaning, tones, references) that support the topic determination of input. Second, deduce a diagnostic REASONING process from premises (i.e., clues, input) that supports the topic determination. Third, determine the overall TOPIC of INPUT as one of CATEGORIES considering CLUES, the REASONING process and the INPUT. MESSAGE: Maintain good hygiene habits to keep illnesses at bay. Regular handwashing, dental care, and a clean living space contribute to your overall health. CATEGORIES: 1. legal_awareness 2. eco_friendly_consumer_choices 3. community_meetings 4. service_to_others 5. career_networking 6. sports_and_physical_activities 7. hygiene 8. public_service 9. voting 10. outdoor_advocacy"

" CLUES: 1. Maintain: implies a continuous effort 2. Good hygiene habits: implies personal care 3. Keep illnesses at bay: implies health 4. Regular handwashing: implies personal hygiene 5. Dental care: implies oral hygiene 6. Clean living space: implies environmental hygiene REASONING: 1. The input emphasizes the importance of maintaining good hygiene habits. 2. The term hygiene is explicitly mentioned. 3. The input implies a continuous effort to keep illnesses at bay. TOPIC: The input is about hygiene."

"Q: Thank you for your reasoning! We need an answer in very strict format so that our automatic systems will be able to process it. It is very important. Please tell what is the final verdict of your expert categorization. Answer in one word. Only use word from the potential topics. Any deviation from these rules will result in system crash. A: The final verdict in single word is: "

" hygiene."

Example COT_2 prompting response of MISTRAL INSTRUCT 0.2 7B model:

"Q: Break down the message Explore new recipes and ensure your diet is well-balanced with a mix of fruits, vegetables, and proteins. Your nutrition is the foundation of good health. and discuss its potential categories from the list: 1. professional_development 2. financial_education 3. nutrition 4. conflict_resolution 5. formal_education 6. empathy_and_understanding. Start by eliminating unlikely options."

A: This message primarily relates to the nutrition category. The message is about exploring new recipes to maintain a balanced diet, emphasizing the importance of fruits, vegetables, and proteins, and stating that nutrition is the foundation of good health. This message is not about professional development, financial education, conflict resolution, formal education, or empathy and understanding. The message is solely focused on nutrition.

"Q: Narrow down to the most appropriate category for the message. State the final verdict in a single word or token."

" A: nutrition."

Example COT_1 prompting response of OPENHERMES MISTRAL 2.5 7B model:

“Q: You are a message expert. You have seen many of them. You are always able to guess what the main topic of a message is. You are always shown a list of possible categories and you always explain why each one of them is or is not related to the original message. You are precise and professional. You do only what you are expected to do. You say no redundant words. You are given a message with following content: ”Explore new recipes and ensure your diet is well-balanced with a mix of fruits, vegetables, and proteins. Your nutrition is the foundation of good health.” This message can be categorized as only one of the following potential topics: 1. professional_development 2. financial_education 3. nutrition 4. conflict_resolution 5. formal_education 6. empathy_and_understanding. A: For each of the possible categories I will think if it could be the correct one for the message:”

“1. professional_development: This message is not about professional development. It is not about improving one’s skills or knowledge for a job or career. 2. financial_education: This message is not about financial education. It is not about learning about money or how to manage it. 3. nutrition: This message is about nutrition. It is about the importance of a well-balanced diet for good health. This is the correct category. 4. conflict_resolution: This message is not about conflict resolution. It is not about how to manage or resolve conflicts. 5. formal_education: This message is not about formal education. It is not about learning in a structured educational setting. 6. empathy_and_understanding: This message is not about empathy and understanding. It is not about learning to understand or care about others. In conclusion, the message is about nutrition.”

“Q: Thank you for your reasoning! We need an answer in very strict format so that our automatic systems will be able to process it. It is very important. Please tell what is the final verdict of your expert categorization. Answer in one word. Only use word from the potential topics. Any deviation from these rules will result in system crash. A: The final verdict in single word is: “

“ Nutrition.”