COGNAC at SemEval-2025 Task 10: Multi-level Narrative Classification with Summarization and Hierarchical Prompting

Azwad Anjum Islam & Mark A. Finlayson
Florida International University
Knight Foundation School of Computing and Information Sciences
11200 SW 8th Street, Miami, FL 33199, USA
{aisla028, markaf}@fiu.edu

Abstract

We present our approach to solving the *Narrative Classification* portion of the *Multilingual Characterization and Extraction of Narratives* SemEval-2025 challenge (Task 10, Subtask 2) for the English language. This task is a multi-label, multi-class document classification task, where the classes were defined via natural language titles, descriptions, short examples, and annotator instructions, with only a few (and sometime no) labeled examples for training. Our approach leverages a text-summarization, binary relevance with zero-shot prompts, and hierarchical prompting using Large Language Models (LLM) to identify the narratives and subnarratives in the provided news articles. Notably, we did not use the labeled examples to train the system. Our approach well outperforms the official baseline and achieves an F_1 score of 0.55 (narratives) and 0.43 (subnarratives), and placed 2nd in the test-set leaderboard at the system submission deadline. We provide an in-depth analysis of the construction and effectiveness of our approach using both open-source (LLaMA 3.1-8B-Instruct) and proprietary (GPT 4o-mini) Large Language Models under different prompting setups.

1 Introduction

Disinformation, misinformation, propaganda, and foreign malign influence (FMI) have become serious problems in the modern information environment. One commonality amongst them is the use of *narrative* to drive their effects. A *narrative* can be defined as a concise, concrete description of a set of events involving a small number of actors, often supporting an evaluative judgment. The ability to automatically identify narratives in textual materials (for example, news or social media) would be of great use to tracking, understanding, and mitigating pernicious influence.

Task 10 at SemEval-2025, *Multilingual Characterization and Extraction of Narratives from Online News* (Piskorski et al., 2025), focuses on automatic identification of different types of narratives and subnarratives, as well as identifying the roles of the relevant entities in news articles. The task is divided into three subtasks—*Entity Framing*, *Narrative Classification*, and *Narrative Extraction*—spanning five languages: Bulgarian, English, Hindi, Portuguese, and Russian. We work on the *Subtask 2: Narrative Classification* for English, where we develop a prompt-based approach to identify narratives and their subtypes in news data.

Subtask 2 defines two domains (*Climate Change* [CC] and *Ukraine-Russia War* [URW]), for which the task creators have defined a set of top-level narratives, each having specific subnarratives. Each news article associated with the domains is labeled with some number of top-level narratives and subnarratives, with no restriction on the number of labels. Each top-level narrative and subnarrative is defined by a title (e.g., *Criticism of Climate Policies*), plus a longer definition, instructions, and zero or more examples. Thus, Subtask 2 is a multi-label, multi-class document classification task.

Our approach has three stages. First, we apply a summarization step that condenses the target document (i.e., a news article) into a uniform length, information-dense representation. Second, we apply class-specific zero-shot prompts using a binary relevance strategy (Zhang et al., 2018) to classify each document as to its top-level narrative category, aggregating results to generate multi-label outputs. Third, we use hierarchical prompting (Liu et al., 2021) to produce subnarrative labels for each narrative found in the articles. We experiment with both open-source (LLaMA 3.1-8B-Instruct) (Meta, 2024) and proprietary models (GPT-4o-mini) (OpenAI, 2024) . Notably, our approach does not use any of the labeled training data for fine-tuning or other model optimizations (barring the experiment comparing zero-shot vs. few-shot setup, where we find that the zero-shot approach performs bet-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1442–1449

ter overall). Our system achieved an F_1 score of 0.55 for narratives and 0.43 for sub-narratives, placing 2nd in the official leaderboard for English (the leading system obtained scores of 0.59 and 0.44, respectively). It is notable that our approach was competitive despite the absence of computationally expensive model training.

The remainder of the paper is structured as follows. We first provide background on the topic of narrative classification and prompt-based solutions in general (§2). We next describe the data and task definition provided by the task organizers (§3). We then elaborate on our methodology and experimental set-up (§4), and report the result from the official submission along with some additional experiments (§5). Finally, we enumerate our contributions and discuss our findings, limitations, and scope for future improvements (§6).

2 Related Work

Multi-label document classification is the task of assigning multiple relevant labels or categories to a text, as opposed to a single label (Tsoumakas and Katakis, 2007). Traditional Machine Learning (ML) and Natural Language Processing (NLP) approaches have developed various methods to tackle this problem. One common approach is *binary relevance*, where the task is decomposed into independent binary classification problems for each label (Zhang et al., 2018). Another is *classifier chains*, which extends binary relevance by linking classifiers in a chain, allowing label predictions to influence one another and capture label dependencies (Read et al., 2011). A third method, *label power-set*, treats each unique combination of labels as a separate class, transforming the problem into a mutually exclusive multi-class classification task (Madjarov et al., 2012).

With the development of large language models (LLMs) (Radford et al., 2019; Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023), a new paradigm for text classification has emerged: instead of training a model specifically for a classification task, it is now possible to *prompt* a pretrained LLM to classify text by describing the task in natural language. This approach has gained widespread popularity as LLMs have shown strong performance in new classification tasks through in-context learning (ICL) with just a few prompt examples (Brown et al., 2020). In scenarios where little to no training data is available, this approach is especially attractive.

Peskine et al. (2023) showed how LLMs can use class definition to produce multi-label classification with zero-shot prompting. This label-by-label prompting in an one-vs-rest manner is simple and allows the model to focus on a binary question each time, potentially improving reliability for each label. This approach is more computationally expensive due to requiring multiple queries for each input. An alternative prompting strategy to increase efficiency is to prompt the LLM to produce all desired labels in one pass. However, this approach increases the classification complexity, which can lead to reduced performance. (Trust and Minghim, 2024; Kostina et al., 2025) Additionally, classification involving multiple classes require more sophisticated prompts and reasoning steps to guide the model, which may also increase format deviation in a model's output, compared to binary classification with simple yes/no format (Kostina et al., 2025).

3 SemEval-24 Task 10 Data

SemEval-2025 Task 10 focuses on analyzing news articles in five languages: Bulgarian, English, Hindi, Portuguese, and Russian and comprises three separate subtasks. We work on subtask 2 in English, which is a multi-label, multi-class narrative classification task. Given a news article and the two-level taxonomy of narrative labels for each domain (where each narrative is subdivided into subnarratives), the task is to assign the article all the appropriate narrative and subnarrative labels. The two domains for this task were: the Ukraine-Russia War (URW) and Climate Change (CC). Each narrative and subnarrative is defined by a title, a short definition, zero or more example statements, and sometimes, additional instructions. For example: the narrative ***Criticism of climate policies*** under the CC domain is defined as: *Statements that question the effectiveness, economic impact, or motives behind climate policies. Example: "It is all because of the decision to switch to electric."* while the subnarrative ***Climate policies are ineffective*** under this narrative is defined as: *Statements suggesting that climate policies fail to achieve their intended environmental goals. Example: "There is absolutely no point in banning straws, it can even have the opposite effect."*

The English training data comprised 399 articles, with 176 from the *Climate Change* domain and 223

from the *Ukraine-Russia War* domain. Additionally, a development dataset of 41 labeled articles (24: *CC*, 17: *URW*) and a test dataset of 101 unlabeled articles (48: *CC*, 53: *URW*) were released by the task organizers. Each article in the training and development dataset is annotated with one or more high-level narrative(s) as well as corresponding finer-grained subnarrative(s). In the case where specific narrative or subnarrative label could not be assigned, the "Other" pseudo-label was used.

4 Approach

Our approach for the Narrative Classification subtask has three steps: (1) summarization that makes the articles more uniform in length and style (§4.1); (2) a set of zero-shot, class-specific LLM prompts to produce binary outputs for each top-level narrative class (§4.2); and (3) a hierarchical prompting technique to sequentially identify subnarrative classes only when the corresponding narrative classes are detected (§4.3).

In this task, the number of class and subclass labels was large compared to the available labeled data. Figure 1 shows the distribution of the number of available labeled samples per subnarrative class. Notably, some subnarrative classes never appear in the English training data. This rendered approaches like supervised learning and fine-tuning problematic for those classes. Therefore, we opted for a prompt-based approach using pre-trained LLMs. For our experiments, we chose one open-source model (LLaMA 3.1-8B-Instruct) and one proprietary model (GPT-4o-mini). It is worth mentioning that, in the English training data, the domain of each input article is indicated in the filename, so we assume knowledge of the domain for all experiments. However, we also conducted an auxiliary experiment that showed LLMs could automatically identify[1] the domain of the input texts in 99% of cases (435 out of 440 articles) across the training and validation datasets.

4.1 Article Summarization

Long-form news articles often contain statements that are not directly related to the main themes of the article. We sometimes found the LLMs to be confused by statements that are either not important, or less important, in the overall context of the article, resulting in false positive labels. To counter this issue, we first experimented **prompt tuning**, adjusting the prompt by instructing the model to base its decisions only on statements pertinent to the main themes of the article, while keeping the input unchanged. We also tried **summarization**, in which we prompted the same LLM used for the classification steps (LLaMA 3.1-8B-Instruct or GPT-4o-mini) to summarize the article into a more concise and information dense form (the prompt is shown in Appendix A).

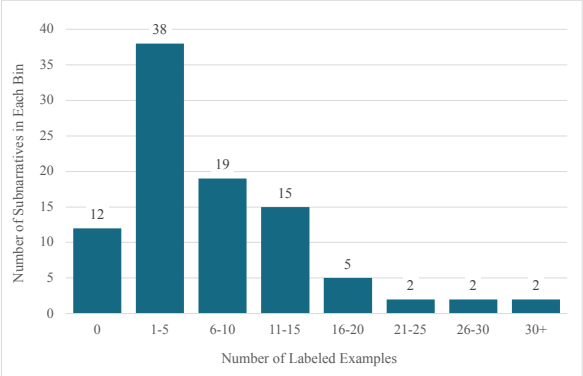

Figure 1: Summary of available English data samples per subnarrative class (including pseudo-labels) in the training data, grouped into bins

Table 1 shows the differences between classification performance on the development data after top-level narrative classification in step 2 (§4.2), depending on whether the texts were unmodified, subjected to prompt tuning, or summarized. We see that summarization results in significant improvement in case of LLaMA, whereas the effect is less pronounced with the more advanced GPT model.

It is worth noting that in the top-level narrative classification, having a high accuracy score is especially important, as misclassifications at this level are guaranteed to produce more errors in subsequent subnarrative classifications. We define accuracy by the percentage of decisions taken by the model that were correct. Correct decisions include assigning correct labels as well as not assigning incorrect labels to articles.

4.2 Binary Relevance with Zero-Shot Prompting

The task is a multi-label multi-class classification problem with more than ten narratives in each domain. Use of a single large prompt to identify all the correct classes in a text risks putting a burden of excessive information on the LLM. It also makes it harder to define the different narratives effec-

[1]We prompted the GPT-4o-mini model with the prompt: "Given the following text text, determine if its content is primarily about *Climate Change* or *Ukraine Russia War*".

tively while keeping the prompt concise and clear. To alleviate this problem, we treated the top-level multi-label classification task as a series of binary classification tasks using the binary relevance technique. For each narrative class, a class-specific prompt was developed using the definition, example statements, and optional annotation instruction provided in the official taxonomy. The prompts were developed following recommended practices of prompt engineering:

Persona: Researchers have found role-play prompting to consistently surpass the standard zero-shot approach across most datasets (Kong et al., 2024; Tseng et al., 2024). We assigned the model in our experiments the role of an "expert narratologist" in the corresponding domain.

Context: We provide relevant context to the model including the definition of narratives in the taxonomy, descriptions of common subnarratives with example statements, and additional instructions when available in the taxonomy.

Clear instructions: We clearly outline the task and provide step-by-step guidelines for the model to follow, which has been shown to improve model response (Wu et al., 2023).

Chain of Thought (CoT): Eliciting a series of intermediate reasoning steps can significantly improve complex reasoning capabilities of LLMs (Wei et al., 2023). In our experiments, we take advantage of zero-shot CoT by prompting the LLM to produce the intermediate reasoning steps.

Output format: We explicitly specify the desired output format in our prompts to avoid output inconsistency and format deviation, which is crucial to ensure accurate parsing of narrative labels generated by the model (Liu et al., 2024).

The template for the narrative classification prompts is given in Appendix B. Table 2 shows the performance improvement achieved with the binary relevance method over using a single prompt.

We also compared zero-shot with few-shot prompting. Few-shot prompting is often favored over zero-shot prompting as the former generally produces more accurate results (Brown et al., 2020). For the narrative-classification task, we experimented with 0-shot, 2-shot, and 4-shot prompting using the two LLMs, while using the summarized articles as input. We randomly selected one (or two) positive example(s) and an equal number of negative example(s) from the training data to produce the 2-shot (or 4-shot) prompts for this experiment. This is the only experiment where we make use of the labeled training data. Table 3 shows the comparative performance across different shot settings. We see that 0-shot prompting achieves better accuracy overall, while the micro F_1 score remains consistent across different shot settings. The difference is again more significant for the LLaMA model compared to the GPT model.

Method	CC		URW	
	Acc.	F_1	Acc.	F_1
LLaMA 3.1-8B-Instruct				
Unmodified	0.66	0.35	0.52	0.44
Prompt Tuning	0.68	0.39	0.50	0.42
Summarization	**0.82**	**0.48**	**0.82**	**0.47**
GPT-4o-mini				
Unmodified	0.88	0.55	0.84	0.57
Prompt Tuning	0.88	0.59	0.86	**0.61**
Summarization	**0.89**	**0.59**	**0.87**	0.57

Table 1: Performance differences in top-level narrative classification on the development dataset

Method	Acc.	F_1
LLaMA 3.1-8B-Instruct		
Single prompt	0.73	0.39
Binary relevance	0.82	**0.47**
GPT-4o-mini		
Single prompt	0.80	0.48
Binary relevance	0.88	**0.58**

Table 2: Performance comparison between single prompt and binary relevance for top-level narrative classification

4.3 Hierarchical Prompting for Subnarrative Detection

Once the top-level narratives had been identified, we used hierarchical prompting to classify subnarratives in each text by using class-specific prompts for each of the narratives. If the LLM classifies a narrative as present in an article in the top-level narrative classification step, the model is then subsequently prompted to identify the appropriate subnarrative(s) in the article with a prompt specific to that particular narrative class. This sequential approach simplifies the subnarrative classification process for the LLM model by providing information about the presence of the top-level narrative, as well as reducing the number of possible classes. Notably, we do not use the binary relevance method in this level considering the comparatively small number of possible classes and computational complexity. The subnarrative classification prompts were developed following the same prompt engi-

Method		Acc.	F_1
LLaMA 3.1-8b-Instruct	0-shot	**0.82**	**0.47**
	2-shot	0.70	**0.47**
	4-shot	0.68	0.46
GPT-4o-mini	0-shot	**0.88**	**0.59**
	2-shot	0.84	**0.59**
	4-shot	0.82	0.57

Table 3: Performance comparison across different shot settings in top-level narrative classification

Top-level classification Strategies	**Samples F_1**	
	Full	**Summarized**
LLaMA 3.1-8B-Instruct		
0-shot	0.38	**0.39**
2-shot	0.27	0.28
4-shot	0.25	0.26
Prompt-tuning	0.23	0.25
GPT-4o-mini		
0-shot	0.52	**0.55**
2-shot	0.51	0.51
4-shot	0.48	0.50
Prompt-tuning	0.51	0.52

Table 4: Performance in subnarrative classification using full and summarized input articles, paired with different top-level narrative classification strategies.

neering practices described in section 4.2. The template for these prompts is given in Appendix C.

For the subnarrative classification, we experimented with both the full articles and their summaries as inputs to the LLM. These two input strategies were paired with multiple top-level narrative classification strategies described previously (i.e., using prompt-tuning on full articles, using 0-shot prompting on summarized articles etc.) to construct different variants of the pipeline. Table 4 shows the results. Interestingly, summarized inputs produced better results even for fine-grained subnarrative classification. This may be attributed to the normalizing effect the summarization step had on the structure and readability of the input articles.

5 Results

For the final submission, we chose the best performing model based on our experiments on the development set, which was GPT-4o-model with zero-shot prompts using summarized articles as input in both narrative and subnarrative classification step. Table 5 shows results from the official evaluation of our methods on the unlabeled test dataset, together with other top performers. The official evaluation measure for ranking was the averaged

Team	**F_1 macro coarse**	**F_1 st. dev. coarse**	**F_1 samples**	**F_1 st. dev. samples**
GATENLP	0.590	0.353	**0.438**	0.333
COGNAC*	0.554	0.400	**0.426**	0.391
INSALyon2	0.513	0.378	**0.406**	0.382
23	0.493	0.392	**0.377**	0.384
NCLteam	0.486	0.363	**0.345**	0.360

Table 5: Official test set results (Top 5). Our team, COGNAC, is marked by an asterisk (*)

samples F_1 score computed for the subnarrative labels. We achieve high F_1 score in both narrative and subnarrative classification, and rank 2nd despite not using the labeled training dataset, or computationally expensive model fine-tuning.

Our experiments showed that summarization significantly enhanced classification performance. We attribute this to summarization making the input more information-dense and uniformly structured. We also found that breaking down a complex multi-label classification problem into a series of simpler binary classifications led to better performance, which is consistent with the observations of many other researchers that LLMs are better at performing simple tasks with a clear goal compared to complex tasks with more sophisticated instructions. These effects were much more pronounced with the smaller model, which is consistent with the assumption that more advanced LLMs are better at handling complex tasks. It also indicates that using a larger, state-of-the-art model may further improve the performance of our approach.

5.1 Error Analysis:

As the training data was not used for for training or fine-tuning the model, we were able to leverage it for the purpose of error analysis. We found that the system exhibited a conservative prediction strategy, favoring under-prediction. At the top-level narrative classification, only 17% of classification errors were false positives, while 83% were false negatives. However, this behavior was intended, and was achieved through instruction-tuning (e.g. via phrases like "...the provided text *explicitly* includes the narrative...", "...such statements are *prominently* present..." etc.. See Appendix B). Due to the overwhelmingly large number of true negative cases for all narrative classes compared to true positives, a less conservative prompt—while reducing some false negatives—causes substantial rise in false positive errors and negatively affects overall performance. Due to this conservative approach, the

LLMs often did not identify narratives when the narratives were only subtly indicated in the text. For example, the phrase "Russia is at war with pure evil" did not trigger a positive identification for the narrative "Praise of Russia". Also, upon manual inspection of some of the reasoning steps produced by the LLMs, we noticed that both LLMs were able to produce reasonable and coherent chain-of-thoughts behind their answers most of the time, but they often fell short in identifying more nuanced clues necessary for complex narrative types. For example, the LLMs frequently failed to distinguish between "Criticism of international entities" and "Criticism of political organizations and figures" subnarratives under the narrative "Criticism of institutions and authorities".

6 Conclusion

In this paper, we proposed a prompt-based approach to the multi-label multi-class narrative classification problem introduced at SemEval-2025 (Task 10, Subtask 2) for the English language. We leveraged text-summarization, binary relevance with Large Language Models (LLMs), and hierarchical prompting technique to label broad narratives as well as fine-grained subnarratives in news articles. We developed zero-shot prompts for each narrative and subnarrative class solely from the provided taxonomy, avoiding the need for training data and expensive model fine-tuning. This approach achieved competitive performance, placing 2nd in the leaderboard.

Despite promising results, there is much room for improvement. Our approach does not consider the possibility that some labels may be more likely to co-occur together. While we take advantage of class-specific prompts in a binary relevance approach, further refinement of these prompts with help of domain experts could help address specific weaknesses in individual narrative classifications. Additionally, we only applied our approach for the English dataset, which leaves its multilingual capability an open question for future studies.

7 Acknowledgements

This work was partially supported by the Defense Advanced Research Projects Agency under contract number HR001121C0186.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, New York, NY. Curran Associates, Inc.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico.

Arina Kostina, Marios D. Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review. *Preprint*, arXiv:2501.08457.

Hui Liu, Danqing Zhang, Bing Yin, and Xiaodan Zhu. 2021. Improving pretrained models for zero-shot multi-label text classification through reinforced label hierarchy reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1062, Online.

Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J. Cai. 2024. "We Need Structured Output": Towards User-centered Constraints on Large Language Model Output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY.

Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. 2012. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104.

Meta. 2024. Llama 3.1 8b instruct.

OpenAI. 2024. Gpt-4o mini.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, New York, NY. Curran Associates, Inc.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding GPT for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063, Singapore.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Ricardo Guimarães, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Jesse Read, Bernhard Pfahringer, Geoffrey Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Paul Trust and Rosane Minghim. 2024. A study on text classification in the age of large language models. *Machine Learning and Knowledge Extraction*, 6(4):2688–2721.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida.

Grigorios Tsoumakas and Ioannis Manousos Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3:1–13.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yang Wu, Yanyan Zhao, Zhongyang Li, Bing Qin, and Kai Xiong. 2023. Improving cross-task generalization with step-by-step instructions. *Preprint*, arXiv:2305.04429.

Min-Ling Zhang, Yukun Li, Xu-Ying Liu, and Xin Geng. 2018. Binary relevance for multi-label learning: An overview. *Frontiers of Computer Science*, 12(2):191–202.

A Summarization Prompt

```
Summarize the following input text and output the summary in 300 words or less. Retain all the main topics, sentiments, and narratives of the text.
```

B Narrative Classification Prompt Template

We used the following template to generate narrative classification prompts from the official taxonomy.

```
Role: You are an expert narratologist skilled in analyzing and identifying narratives within text, particularly in the domain of <domain>.
Determine whether the provided text explicitly includes the narrative <narrative>

Definition of the Narrative:
The narrative <narrative> is defined by <definition from taxonomy>.
Example of statement that aligns with this narrative: <example from taxonomy>
Common Themes within this Narrative may include: <list of subnarratives, with definition and example from taxonomy>
Or other statements supporting <narrative>.

Follow these guidelines: Read the provided text carefully.
Find if there are any statements the strongly support the narrative <narrative>.
Answer "Yes" if such statements are prominently present.
Answer "No" if statements supporting the narrative <narrative> are not prominently present.
Explain your reasoning for the decision, referencing specific statements from the text.

Output format:
The first line should be a single word, either "Yes" or "No", depending on your decision. The second line should contain your reasoning for your decision, in a single paragraph.

Input:
```

C Subnarrative Classification Prompt Template

We used the following template to generate narrative classification prompts from the official taxonomy.

```
Role: You are an expert narratologist skilled in analyzing and identifying narratives within text, particularly in the domain of <domain>.
Your task is to analyze a given text that contains the narrative <top-level narrative> and identify what specific subtype of the narrative is present in the text.

The narrative <top-level narrative> may have the following subtypes:
<subnarrative 1>, which is defined by <definition from taxonomy>.
Example of statements supporting this subtype: <example from taxonomy>
<subnarrative 2>, which is defined by <definition from taxonomy>.
Example of statements supporting this subtype: <example from taxonomy>
.
.
.
'Other' which is defined by the absence of the previously mentioned specified subtypes.

Task:
Read the input text.
Decide which one of the specified subtypes of <top-level narrative> are present in the text.
Carefully consider the distinctions between their definitions and choose the best one.
If you cannot choose one best answer, it is possible to answer multiple subtypes.
Answer "Other" if you don't find any of the specified subtypes.
Also explain your reasoning.

Output format:
First line of output should only be the name of the subtype. If your answer is more than one subtypes, they should be separated by commas(,).
Second line of output should be the reasoning for your answer in a single paragraph.

Input:
```