GPLSICORTEX at SemEval-2025 Task 10: Leveraging Intentions for Generating Narrative Extractions

Iván Martínez-Murillo[1], María Miró Maestre[1], Aitana Morote Martínez[1], Snorre Ralund[2], Elena Lloret[1], Paloma Moreda[1], Armando Suárez Cueto[1],
[1]University of Alicante, Carr. de San Vicente del Raspeig, s/n,
San Vicente del Raspeig, Alicante, Spain, 03690
[2]University of Copenhagen, Center of Social Data Science,
Øster Farimagsgade 5, 1353 Copenhagen, Denmark

Abstract

This paper describes our approach to address the SemEval-2025 Task 10 subtask 3 for the English language, which is focused on narrative extraction given news articles with a dominant narrative. We design an external knowledge injection approach to fine-tune a Flan-T5 model so the generated narrative explanations are for the provided dominant narrative in each text. We also incorporate pragmatic information in the form of communicative intentions, using them as external knowledge to assist the model. This ensures that the generated texts align more closely with the intended explanations and effectively convey the expected meaning. The results show that our approach ranks 3rd in the task leaderboard (0.7428 in Macro-F1) with concise and effective news explanations. The analyses highlight the importance of adding pragmatic information when training systems to generate adequate narrative extractions.

1 Introduction

Understanding natural language goes beyond recognizing objects and their relations within a sentence or a news article. The way information is presented—or omitted—is shaped by the author's communicative intentions and implicit biases. To fully grasp the narrative conveyed by a text, one must move beyond a surface-level reading and incorporate assumptions about the author's intent, as well as general knowledge about the topic.

With the rise of modern misinformation, automated approaches have become crucial in combating information warfare. However, these systems come with their own biases, making explainability a key factor in building trust. SemEval-2025 Task 10 addresses this challenge by combining the identification and explanation of narratives in text. Our work focuses on Subtask 3, which involves generating narrative extractions given a news article. Large Language Models (LLMs) have been shown to encode general knowledge (Ju et al., 2024; Wang et al., 2023), and analytical capabilities (Chang et al., 2023) to identify implicit narratives of a text. Additionally, they hold promise for explainability, as they can be prompted to generate explanations (Roy et al., 2023; Yang et al., 2023). However, their explanations still fall short of human-level quality (Di Bonaventura et al., 2024).

Our approach enhances LLM-generated explanations as a form of controllable abstractive summarization (He et al., 2020) using knowledge injection and standard fine-tuning to align the model's outputs with the required task format. A key contribution of our work is leveraging a curated dataset with labeled communicative intentions to assist the model's analytical capabilities based on Speech Act Theory (Austin, 1962). We hypothesize that injecting this knowledge about intentions can control the model's focus and improve explanation quality. Our approach ranked 3rd in the competition.

2 Background

2.1 Task Description

SemEval-2025 Task 10, "Multilingual Characterization and Extraction of Narratives from Online News" (Piskorski et al., 2025), addresses the identification and analysis of manipulative and harmful disinformation in news articles. The task comprises three subtasks, applied to news articles available in five languages: Bulgarian, English, Hindi, Portuguese, and Russian.

Our submission pertains to the English-language configuration of Subtask 3, which focuses on generating a free-text explanation of given news articles based on the predominant narrative they convey. The provided data for generating the explanation consists of two inputs: the dominant narrative embodying the text intention, which is composed of two labels (dominant narrative and dominant subnarrative) extracted from a narrative taxonomy, and

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 575–583
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

the news articles, containing the ground for the narrative tools making up the given intention.

The dataset for Subtask 3 in English is composed of three sets: (1) Training set and (2) Validation set, containing 203 and 30 articles respectively, each accompanied by its corresponding narrative, subnarrative, and expected output annotations; and (3) Test set, consisting of 68 articles, annotated only with their narrative and subnarrative (Stefanovitch et al., 2025).

2.2 Narrative Extraction Task

Task 10 Subtask 3 at SemEval 2025 focuses on generating narrative extractions from a given news article embedded within a text. Narrative extraction refers to the use of computational techniques to identify, link, and visualize narrative elements from textual sources (Santana et al., 2023). This process involves several key steps: Information Retrieval, which aids in locating relevant information; Text Summarization, which condenses and integrates information pertinent to the narrative; Natural Language Processing, which identifies, extracts, and connects narrative components; and Natural Language Generation, which transforms structured data into coherent textual explanations.

As a text-to-text generation task, narrative extraction is closely related to summarization, but with a distinctive analytical focus on the author's intent and persuasive strategies—aligning it with the concept of controllable text summarization (He et al., 2020; Urlana et al., 2024). A wide range of techniques have been employed in summarization tasks, including statistical machine learning, unsupervised methods, supervised deep learning, and fine-tuning of pretrained language models (Zhang et al., 2024; Urlana et al., 2024).

The advent of instruction-tuned large language models has significantly enhanced the flexibility of controllable summarization, enabling fine-grained control over style, length, and output format through prompt engineering, few-shot learning, and fine-tuning strategies (Liang et al., 2024).

In this context, our approach introduces a novel combination of intent modeling—to guide the generation of narrative extractions—and example-based fine-tuning—to align the system with the specific requirements of the task, including the desired style and format of the output.

2.3 Communicative Intentions

Regardless of its genre, register, or topic, every text inherently carries a communicative intention. Beyond its significance in linguistic research (Austin, 1962; Searle, 1969; Sperber and Wilson, 1986; Bach, 2012), this pragmatic element has also attracted attention in the Natural Language Processing (NLP) community. Advances in NLP now enable the automatic analysis of discourse-related phenomena, making it possible to extract pragmatic information that was previously unfeasible to study due to the complex inferential processes involved in detecting intentions (Mahowald et al., 2024).

Within the existing approaches to the study of intentions, the Speech Act Theory (SAT) (Austin, 1962; Searle, 1969) has been recently applied to many well-established NLP tasks such as question answering (Mirzaei et al., 2023) or text summarization (Mu et al., 2023). Similarly, narrative texts have benefited from the analysis of speech acts (i.e., the actions performed by speakers through their utterances (Yule, 2022)), which helps to better understand the intentions behind narrating events (Kampf, 2021; Borchmann, 2024; Obasi, 2024).

Building on these previous works, we explore the use of intentions for the narrative extraction task to assess whether they help generative models produce more adequate explanations.

3 System Overview

We propose a three-step approach for addressing the narrative extraction task[1]. Our methodology begins by retrieving the narrative intention of the input article, as each article in the dataset exhibits a distinct underlying intention. Identifying these intentions is crucial for shaping the generated explanations. Next, we employ a prompt engineering strategy to construct an input prompt by combining the extracted intention with the corresponding dataset entry. Finally, we inject the knowledge by fine-tuning instruction-tuned models to generate the corresponding explanations. The overall approach is illustrated in Figure 1 and will be described in detail in the following sections.

3.1 Intention Extraction

In this stage, we assumed that the "dominant narrative" assigned to each news article shows great

[1] All code is publicly available on `https://github.com/imm106/teamgplsi-task10`.

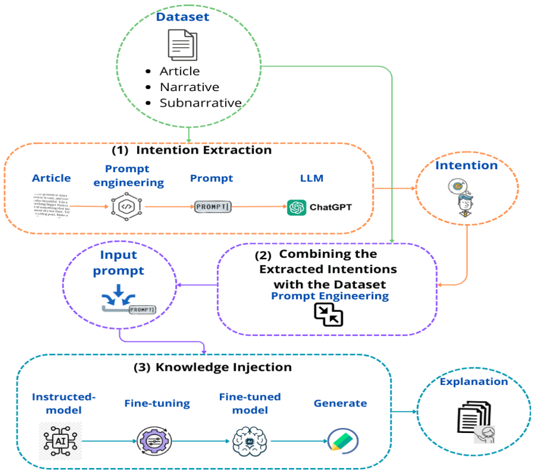

Figure 1: GPLSICORTEX approach to address SemEval 2025 Task 10 Subtask 3.

similarity to what can be considered the text's primary communicative intention. Examples of such intentions include **questioning** a political party, **criticizing** a legal measure or **praising** a political candidate. Therefore, we believed that identifying the intentions conveyed in the narratives would benefit the model's training, enabling the generation of explanations that align with the essential content of the original texts.

To do so, we adapted the work of Maestre et al. (2025), who created a communicative intention annotation scheme for the Spanish language based on two textual dimensions: the intention of the individual segments that shape a message and the global intention of the whole message. More concretely, we selected the 13 global intention categories they established within the annotation scheme and translated them into English, as they showed clear similarities with the intentions observed in the texts provided for the task. The 13 intention categories are: "informative", "personal opinion", "suggestion", "command", "request", "question", "threat", "promise", "praise", "criticism", "emotional", "desire", and "sarcasm / joke".

After establishing the set of intentions to be identified in the narrative texts, we translated the prompt Maestre et al. (2025) used to classify intentions into English so we could classify the global intentions expressed in the narratives. The specific prompt used for this classification task is provided in Appendix A. Based on their findings, we selected GPT-4o-mini, as it was identified as the most effective LLM for automatic intention classification in Spanish. Upon completing the annotation process, we obtained intention tags for each text to enhance the Natural Language Generation (NLG) system. The statistical results of the intention classification are presented in Appendix B.

3.2 Combining the Extracted Intentions with the Dataset

To integrate the extracted knowledge—represented as communicative intentions—into a single input for the model alongside the corresponding dataset entry, we employed a prompt engineering strategy. This approach combines role prompting (Gao, 2023) with the inclusion of control tokens (Li et al., 2022) to effectively inject relevant features. Specifically, we guide the model by explicitly instructing it to act as a text explainer, with the objective of providing explanations of the given articles.

To structure the input, we concatenate key features —article, narrative, subnarrative, and intention— using control tokens to clearly delineate each component (e.g., #Article article, #Theme corresponding narrative, #Intent corresponding intention). A sample prompt is given in Appendix C.

3.3 Knowledge Injection

To incorporate the extracted knowledge, we fine-tuned various instruction-based models using our constructed prompts. Specifically, we evaluated different sizes of LLaMA (Dubey et al., 2024) and Flan-T5 (Chung et al., 2024) models, both of which have demonstrated strong performance in NLG tasks. LLaMA is an LLM that excels in various NLG applications, including text summarization (Bogireddy and Dasari, 2024), a task closely related to our narrative extraction objective. In our experiments, we utilized LLaMA-3.2 and evaluated two model sizes: 1B and 3B parameters. Flan-T5 is an instruction-tuned model designed for multi-task generalization, allowing it to tackle tasks beyond its original training (Zeyad and Biradar, 2024). For our experiments, we tested Flan-T5 Large (780M parameters) and Flan-T5 Base (250M parameters).

To fine-tune these models effectively, we adopted two distinct strategies:

- **Flan-T5 Models**: We performed a full fine-tuning to make the model learn the desired output format.

- **LLaMA Models**:We utilized the *Low-Rank Adaptation (LoRA)* technique (Hu et al., 2021), which introduces trainable low-rank matrices into specific layers instead of updating all model parameters. This method significantly reduces memory and computational costs, making it feasible to fine-tune large-scale models like LLaMA efficiently.

4 Experimental Setup

In this section, we outline the experimental setup utilized in our experiments.

4.1 Dataset Split

To monitor potential overfitting during fine-tuning, we split the training set into two subsets: 90% for training and 10% for development. This resulted in 172 articles with their respective annotations for training and 31 articles for development.

The validation and test splits provided by the task remained unchanged, consisting of 30 articles for validation and 68 for testing. This strategy ensured a reliable assessment of our models' generalization performance while preventing overfitting.

4.2 Fine-tuning & Hyperparameters

To implement the different fine-tuning strategies, we used various Python libraries. Specifically, for full fine-tuning of the Flan-T5 models, we utilized the Transformers library (Wolf et al., 2020). For the LLaMA models, we employed the PEFT library (Mangrulkar et al., 2022) to apply the LoRA technique and the TRL library (von Werra et al., 2020) to manage the training process. The parameters used for the LoRA configuration are presented in Table 1.

Parameters	Values
Rank	8, 16, 24 and 32
Target modules	["q_proj", "k_proj", "v_proj", "o_proj"]
Alpha	8, 64 and 128
Dropout	0.00, 0.01 and 0.05
Bias	none
Task type	CAUSAL_LM

Table 1: LORA configuration parameters setup.

Furthermore, we conducted hyperparameter tuning to determine the optimal fine-tuning configuration. Table 2 presents the search configuration used in this process.

Parameters	Values
Train epochs	2, 3, or 4
Learning rate	$1e-4$, $2e-4$, $3e-4$, or $4e-5$
Weight decay	0, 0.1, or 0.2
Optimizer	adamw_torch_fused

Table 2: Hyperparameter tuning.

All the experiments were conducted on a single NVIDIA A100 GPU.

4.3 Evaluation Metrics

To assess our system's performance, we utilized BERTScore (Zhang et al., 2020), the same metric employed in the shared task evaluation. BERTScore measures the similarity between reference and candidate texts using contextual embeddings from a pre-trained BERT model.

Additionally, we conducted a shallow manual analysis of the generated outputs to further evaluate the quality of our models' predictions.

5 Results

In this Section, we report the obtained results through our experimentation, and our official results in the SemEval 2025 Task 10 Subtask 3.

5.1 Development

During the training phase, we conducted a series of experiments using instructed models and various hyperparameter configurations. The results for each model, obtained using the optimal hyperparameter settings, are presented in Table 3.

System	Precision	Recall	Macro-F1
Llama-3.2-1B	0.70988	0.70928	0.70937
Llama-3.2-3B	0.72093	0.72369	0.72200
Flan-T5 Base	**0.78016**	0.71562	0.74613
Flan-T5 Large	0.76931	**0.73429**	**0.75115**

Table 3: BERTScore results on the development set for the Subtask 3 in English.

As observed, Flan-T5 outperforms LLaMA-3.2 for this task. While LLaMA maintains a stable balance between precision and recall, its overall scores remain lower than those achieved by the Flan-T5 models. Among the Flan-T5 configurations, the base model attains higher precision; however, its recall is lower compared to the larger configuration. Consequently, the larger Flan-T5 model achieves the highest overall performance. A manual analysis of the generated explanations reveals that LLaMA tends to produce explanations split into two or three sentences, delving into longer outputs. In contrast, Flan-T5 generates a single, concise sentence that effectively summarizes the entire article.

Therefore, we selected the Flan-T5 Large model as the basis for our approach in the competition.

5.2 Official Test Leaderboard

We finally submitted the Flan-T5 Large model enhanced with the extracted intentions and fine-tuned to address this task. Table 4 shows the official test leaderboard for subtask 3 in English.

	System	Precision	Recall	Macro-F1
1	KyuHyunChoi	0.76686	0.73517	0.75040
2	WordWiz	0.75464	0.73705	0.74551
3	**GPLSICORTEX**	**0.75375**	**0.73274**	**0.74280**
4	TechSSN	0.73886	0.74568	0.74203
5	NarrativeNexus	0.71991	0.74267	0.73085
...	...	...	...	...
14	Baseline	0.65144	0.68344	0.66690

Table 4: Official Results of SemEval Task 10 Subtask 3 with the BERTScore metric.

Our approach secured third place out of 14 participants in the competition, demonstrating strong performance. Specifically, we achieved a Macro-F1 of BERTScore of 0.74280, only 0.0076 lower than the top-performing method, highlighting the competitiveness of our model. Our results suggest that consistent with our validation set findings, the BERTScore metric reflects high precision, indicating that our model generates highly accurate predictions, although the recall rate is slightly lower.

6 Analysis of the Efficacy of Injecting the Intentions

In this section, we aim to demonstrate that incorporating communicative intentions extracted from the text enhances the model's performance.

To validate this, we fine-tuned the best-performing model, FLAN-T5 in its larger configuration, without integrating the extracted intentions. We then compared its performance against the results obtained when intentions were included. Tables 5 and 6 present the classification performance on the validation and test sets, respectively, using the BERTScore metric.

Intent	Learning Rate	Precision	Recall	Macro-F1
Yes	$1e-4$	**0.76931**	**0.73429**	**0.75115**
No	$1e-4$	0.76055	0.71082	0.73460
Yes	$4e-5$	**0.77681**	**0.72630**	**0.75050**
No	$4e-5$	0.78080	0.71188	0.74445

Table 5: Analysis of the intentions on the validation set with BERTScore metric.

Results show that the classification performance for the models enhanced with intentions is consistently higher than for the models without intent across both the validation and test sets. The intentions help to achieve higher precision, recall, and Macro-F1 scores, indicating that those models are more confident and accurate in generating the explanations of the articles.

Intent	Learning Rate	Precision	Recall	Macro-F1
Yes	$1e-4$	**0.75375**	**0.73274**	**0.74280**
No	$1e-4$	0.73905	0.71740	0.72780
Yes	$4e-5$	**0.76754**	**0.73464**	**0.75040**
No	$4e-5$	0.76033	0.72086	0.73961

Table 6: Analysis of the intentions on the test set with BERTScore metric.

7 Post-Competition

After the conclusion of the competition, we entered a phase where we could systematically evaluate our approaches using the test set.

During this process, we found it particularly insightful to explore the impact of different hyperparameter configurations on our best-performing approach. Specifically, during the competition, we observed that certain hyperparameter settings yielded higher precision on the BERTScore metric in the validation set, besides at the cost of reduced recall compared to the configuration used in our final submission, which was more balanced. This effect was particularly pronounced when adjusting the learning rate.

Table 7 presents the results obtained with the fine-tuned FLAN-T5 model, enhanced with the communicative intentions, across different learning rate configurations.

Learning Rate	Precision	Recall	Macro-F1
$1e-4$ **(Official)**	0.76931	**0.73429**	**0.75115**
$4e-5$	**0.77681**	0.72630	0.75050
$3e-4$	0.75579	0.72457	0.73960
$2e-4$	0.75533	0.71551	0.73463

Table 7: Results of the hyperparameter analysis on the validation set on the post-competition period.

As observed, the model trained with a learning rate of $1e-4$ achieves the best overall performance. However, a learning rate of $4e-5$ leads to higher precision while sacrificing recall, resulting in a BERTScore similar to that of the best-performing configuration. This suggested us that maybe the model configuration with $4e-5$ as the learning rate could perform well on the test set. We also generated the explanations for the test set with all the learning rates configurations. Table 8 shows the results on the test sets for our approaches.

The fine-tuned approach with a learning rate of $4e-5$ outperforms our official competition sub-

Learning Rate	Precision	Recall	Macro-F1
$1e-4$ **(Official)**	0.75375	0.73274	0.74280
$4e-5$	**0.76754**	**0.73464**	**0.75040**
$3e-4$	0.74715	0.73234	0.73935
$2e-4$	0.75064	0.73317	0.74143

Table 8: Results of the hyperparameter analysis on the test set on the post-competition period.

mission, achieving a macro-F1 score of 0.75040 on the BERTScore metric. Notably, this configuration surpasses our official results not only in precision but also in recall. This configuration would have secured us 1st place in the competition, matching the score of the top-performing team.

8 Conclusions

In this paper, we present our approach for the SemEval-2025 Task 10 Subtask 3 in English, focused on the generation of explanations of articles containing disinformation. Our method provides narrative intention knowledge to the model within a fine-tuning and hyperparameter-tuning process.

Our submission involves the fine-tuned Flan-T5 Large model, with which we ranked 3rd in the competition, achieving a BERTScore F1 of 0.74280, only 0.0076 lower than the top-performing method.

The results show the importance of intention analysis in exposing misinformation initiatives. The selected 13 global intention categories extracted from the scheme of Maestre et al. (2025) have been proved to be beneficial for the task, as shown in Section 6.

In addition, this work has provided an analysis to determine the best-performing model configuration for the task. Flan-T5 models have demonstrated satisfactory results.

To effectively detect articles that disseminate disinformation, identifying the author's intent is crucial. However, our approach can be further enhanced by integrating additional perspectives. For instance, incorporating a propaganda strategy taxonomy could help identify specific linguistic tools employed by the author to propagate a particular narrative. Furthermore, advanced techniques such as Few-Shot Prompting with LLaMA 3.1, sentiment analysis of dominant narratives, and the Chain of Thought reasoning framework could be utilized to improve the accuracy and depth of the analysis.

Acknowledgments

This research is part of the R&D projects "CORTEX: Conscious Text Generation" (PID2021-123956OB-I00), funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"; "CLEAR.TEXT: Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities" (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and "European Union NextGenerationEU/PRTR"; and project "NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/021)" funded by the Generalitat Valenciana. Moreover, it has been also partially funded by the Ministry of Economic Affairs and Digital Transformation and "European Union NextGenerationEU/PRTR" through the "ILENIA" project (grant number 2022/TL22/00215337) and "VIVES" subproject (grant number 2022/TL22/00215334).

References

John Langshaw Austin. 1962. *How to Do Things with Words*. Oxford at the Clarendon Press.

Kent Bach. 2012. Saying, meaning, and implicating. In *The Cambridge Handbook of Pragmatics*, pages 47–68. Cambridge University Press, Cambridge.

Srinivasa Rao Bogireddy and Nagaraju Dasari. 2024. Comparative analysis of chatgpt-4 and llama: Performance evaluation on text summarization, data analysis, and question answering. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

Simon Borchmann. 2024. Headlines as illocutionary subacts: The genre-specificity of headlines. *Journal of Pragmatics*, 220:73–99.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A Survey on Evaluation of Large Language Models. *arXiv preprint*. ArXiv:2307.03109 [cs].

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara McGillivray. 2024. Is Explanation All You Need? An Expert Survey on LLM-generated Explanations for Abusive Language Detection. In *Tenth Italian*

Conference on Computational Linguistics (CLiC-it 2024). Pisa, Italy.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Andrew Gao. 2023. Prompt engineering for large language models. *Available at SSRN 4504303*.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. CTRLsum: Towards Generic Controllable Text Summarization. *arXiv preprint*. ArXiv:2012.04281 [cs].

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How Large Language Models Encode Context Knowledge? A Layer-Wise Probing Study. *arXiv preprint*. ArXiv:2402.16061 [cs].

Zohar Kampf. 2021. Political speech acts in contrast: The case of calls to condemn in news interviews. *Journal of Pragmatics*, 180:203–218.

Zihao Li, Matthew Shardlow, and Saeed Hassan. 2022. An investigation into the effect of control tokens on text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 154–165.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable Text Generation for Large Language Models: A Survey. *arXiv preprint*. ArXiv:2408.12599 [cs].

María Miró Maestre, Ernesto L. Estevanell-Valladares, Robiert Sepúlveda-Torres, and Armando Suárez Cueto. 2025. Enhancing pragmatic processing: A two-dimension approach to detecting intentions in spanish. *Procesamiento del Lenguaje Natural*, 74. Accepted, awaiting publication.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Maryam Sadat Mirzaei, Kourosh Meshgi, and Satoshi Sekine. 2023. What is the real intention behind this question? Dataset collection and intention classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13606–13622, Toronto, Canada. Association for Computational Linguistics.

Fangwen Mu, Xiao Chen, Lin Shi, Song Wang, and Qing Wang. 2023. Developer-intent driven code comment generation. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 768–780.

Jane Chinelo Obasi. 2024. Pragmatic analysis of news reports on coronavirus from selected national and international newspapers. *Howard Journal of Communications*, 35(4):429–454.

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.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artificial Intelligence Review*, 56(8):8393–8435.

John R Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*, volume 626. Cambridge University Press.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Citeseer.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, Purificação Silvano, Shivam Sharma, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari Gonçalves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub

Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, European Commission Joint Research Centre, Ispra (Italy).

Ashok Urlana, Pruthwik Mishra, Tathagato Roy, and Rahul Mishra. 2024. Controllable Text Summarization: Unraveling Challenges, Approaches, and Prospects – A Survey. *arXiv preprint*. ArXiv:2311.09212 [cs].

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Cunxiang Wang, Sirui Cheng, Zhikun Xu, Bowen Ding, Yidong Wang, and Yue Zhang. 2023. Evaluating Open Question Answering Evaluation. *CoRR*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable Hate Speech Detection with Step-by-Step Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.

George Yule. 2022. *The study of language*. Cambridge university press.

Abdulrahman Mohsen Ahmed Zeyad and Arun Biradar. 2024. Advancements in the efficacy of flan-t5 for abstractive text summarization: a multi-dataset evaluation using rouge and bertscore. In *2024 International Conference on Advancements in Power, Communication and Intelligent Systems (APCI)*, pages 1–5. IEEE.

Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. 2024. A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models. *arXiv preprint*. ArXiv:2406.11289 [cs].

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Prompt to Extract the Communicative Intentions

In this section, we provide a sample prompt used to query GPT-4o for extracting the communicative intention of a text, following the approach of Maestre et al. (2025).

From now on you're going to classify the global communicative intention of the text shown below. The text intention must be one of the following 13 categories: "informative", "personal opinion", "praise", "criticism", "desire", "request", "question", "command", "suggestion", "sarcasm / joke", "promise", "threat" or "emotional". I want your answer to be: The global intention of the text is: Text:

B Analysis of Extracted Intentions

We analyzed the distribution of extracted intentions across each dataset and found notable patterns. Figures 2, 3, and 4 illustrate these distributions. Notably, criticism emerged as the most prevalent intention across all sets. This finding aligns with the dominant narrative of the articles, which exhibited a strong inclination toward critical perspectives. Consequently, this correlation serves as validation of the accuracy of our methodology. Additionally, the second and third most frequent intentions were informative and personal opinion. This pattern can be attributed to the nature of the texts, which are news articles incorporating propaganda techniques. As such, these articles are expected to convey information, often accompanied by subjective points of view.

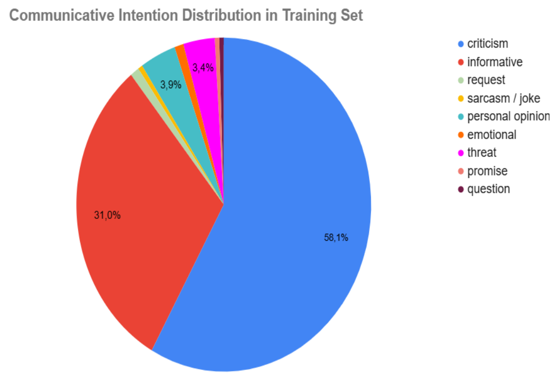

Figure 2: Analysis of the distribution of intentions in the training articles.

One factor that may have influenced our results is the greater diversity of intentions present in the training dataset compared to the validation and test

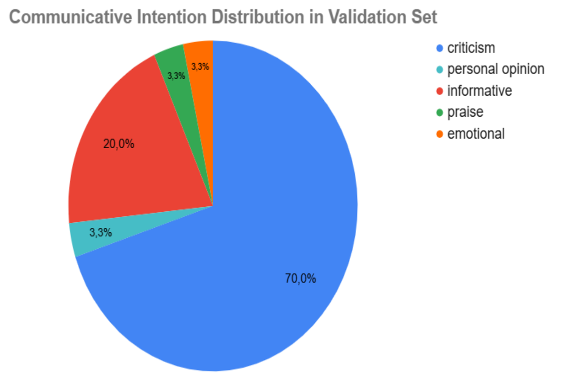

Figure 3: Analysis of the distribution of intentions in the validation articles.

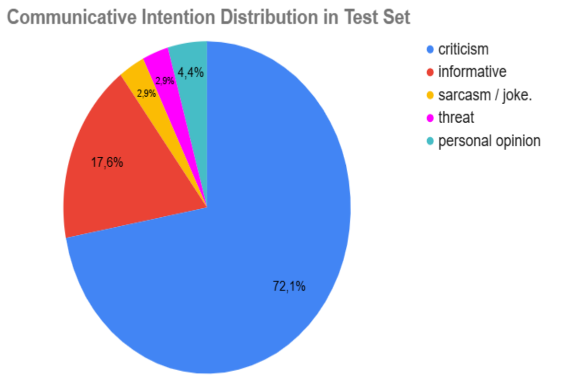

Figure 4: Analysis of the distribution of intentions in the test articles.

sets. However, despite this variation, there was only one instance in which an intention—praise—was detected in the validation set but was absent in the training split.

C Prompt to Combine the Communicative Intention with the Dataset

In this section, we present a sample prompt designed to integrate the extracted intention of the articles with the dataset.

```
From now you will act as a text explainer.
Your task is to generate an explanation
for the following article.

### Article:
This is a sample text. The news article
use to be longer than this one.

### Theme:
Ukraine is the aggressor

### Intent:
Criticism

### Explanation:
```