psy detectives at SemEval-2026 Task 10: PsyCoMark - Psycholinguistic Conspiracy Marker Extraction and Detection

Roxana Carabaş[1] **Anamaria Persida Nacu**[1] **Isac Lucian-Constantin**[1] **Daniela Gîfu**[2,3]
[1]Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi, Romania
[2]Institute of Computer Science, Romanian Academy – Iaşi Branch, Romania
[3]Academy of Romanian Scientists, Romania
carabasroxana@yahoo.com, anamaria_persida@yahoo.com
daniela.gifu@iit.academiaromana-is.ro

Abstract

Psycholinguistic markers provide interpretable signals for identifying conspiratorial reasoning in online discourse. SemEval-2026 Task 10 (PsyCoMark) couples document-level conspiracy labels with span-level annotations across six marker categories, enabling evaluation of both predictive accuracy and evidence alignment. We introduce a set of reproducible baselines combining (i) a marker-derived lexicon and LIWC-style ratio features extracted from gold spans, (ii) supervised transformer models (RoBERTa/DeBERTa) fine-tuned for binary conspiracy detection under optimized hyperparameter regimes, and (iii) a zero-shot TinyLlama-1.1B baseline for the full three-way classification setting. We additionally test a heuristic relabeling strategy based on marker distributions, which does not improve downstream performance. On the official development split, `microsoft/deberta-v3-large` achieves the highest weighted F1 (0.8339) and reaches 0.75 on the competition test set. Results show that transformer-based models remain strong baselines for conspiracy detection, while psycholinguistic markers offer interpretable, human-aligned evidence signals. These baselines establish a controlled reference point for future work integrating marker extraction and supervised three-way classification in psycholinguistically grounded moderation pipelines.

1 Introduction

Conspiracy-related discourse circulates widely across online communities and is closely associated with misinformation propagation, social polarization, and declining institutional trust (Byford, 2011; Douglas et al., 2017). Digital platforms facilitate the rapid diffusion of speculative or hostile narratives and the formation of fringe communities (Spohr, 2017; Zeng and Kaye, 2022), a phenomenon documented extensively in recent analyses of conspiratorial content on TikTok (Corso et al., 2025). In parallel, research in psycholinguistics and discourse studies has shown that communicative uncertainty (Vlăduţescu et al., 2014), multidimensional semantic structures (Gîfu and Cristea, 2012), and persuasive strategies in political language (Gîfu, 2013; Gîfu and Cristea, 2013) are systematically reflected in linguistic patterns. These perspectives align with work on diachronic semantic variation (Gîfu, 2016) and with applied studies on detecting propaganda techniques in news (Ermurachi and Gîfu, 2020), suggesting that linguistic markers can serve as robust indicators of conspiratorial reasoning.

Automatic conspiracy detection is therefore essential for large-scale monitoring and moderation. However, purely predictive models are frequently criticized for limited interpretability, especially when the linguistic evidence driving a decision is not explicit. SemEval-2026 Task 10 (PsyCoMark) (Samory et al., 2026) directly addresses this limitation by coupling document-level conspiracy labels with span-level psycholinguistic marker annotations across six interpretable categories (Actor, Action, Victim, Evidence, Threat, Effect). This design enables the evaluation of both predictive performance and evidence alignment, offering a controlled environment for studying how conspiratorial reasoning is linguistically signaled.

This setting raises a legitimate research question: *How can interpretable psycholinguistic markers be integrated into a robust conspiracy-detection system such that the resulting model is accurate, reproducible, and transparent about the linguistic evidence it relies on?* To address this question, this paper introduces a suite of reproducible baselines tailored to the shared-task context:

- We construct marker-informed lexical resources and compute LIWC-style ratio features derived from gold span annotations, pro-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1–7

viding transparent psycholinguistic representations of documents.

- We establish supervised transformer baselines (RoBERTa/DeBERTa) for binary Yes/No conspiracy detection, using targeted hyperparameter search for small/medium models and efficient step-bounded configurations for large models (Akiba et al., 2019).
- We evaluate a heuristic relabeling strategy based on marker distributions to reduce label ambiguity; empirical results show that it does not improve downstream performance.
- We include a TinyLlama-1.1B zero-shot baseline for the full three-way classification setting (Yes/No/Can't tell), highlighting prompt sensitivity (Zhuo et al., 2024) and reproducibility constraints (Renze, 2024).

Together, these contributions provide a transparent and reproducible foundation for future work integrating psycholinguistic markers into conspiracy-detection pipelines, strengthening the connection between interpretability and performance in computational analyses of online discourse. Our code is released for reproducibility: `https://github.com/carabasroxana/psy_detectives`.

2 Related Work

Research on online conspiracy theories spans psychology, linguistics, computational social science, and natural language processing. Psychological studies have examined the cognitive and social factors that predispose individuals to conspiratorial thinking, highlighting the role of uncertainty, perceived threat, and distrust in institutions (Douglas et al., 2017). These findings align with broader work on communicative uncertainty and discourse ambiguity (Vlăduţescu et al., 2014), which emphasizes how linguistic cues can signal epistemic instability and facilitate the uptake of speculative narratives.

From a linguistic perspective, conspiracy theories exhibit distinctive semantic and pragmatic patterns. Prior work has shown that conspiratorial narratives often rely on implicit causal chains, vague agents, and emotionally charged framing (Fong et al., 2021). Studies on political and persuasive discourse further demonstrate that multidimensional semantic structures and rhetorical strategies can be systematically analyzed and computationally modeled (Gîfu and Cristea, 2012; Gîfu, 2013; Gîfu and Cristea, 2013). Diachronic analyses of lexical semantics also reveal how meaning shifts across contexts and time (Gîfu, 2016), a phenomenon relevant for understanding how conspiratorial language evolves.

In computational settings, conspiracy detection has mainly been framed as document- or post-level classification across news, social media, and online forums (Phillips et al., 2022; Miani et al., 2022). Early approaches relied on lexical, stylistic, and psycholinguistic features, whereas more recent work has shifted toward contextual encoders and LLM-assisted baselines, typically improving predictive performance while making the underlying evidence less explicit. Multimodal platforms such as TikTok have only recently begun to receive systematic attention. A notable example is the large-scale quantitative study by Corso et al. (Corso et al., 2025), which analyzes 1.5M TikTok videos to estimate the prevalence of conspiratorial content and evaluate open-weight LLMs for detecting conspiracy narratives. Their findings—particularly the observation that LLMs achieve high precision but remain sensitive to prompt design—directly motivate our exploration of zero-shot baselines in PsyCoMark. Within this broader literature, PsyCoMark is distinctive because it jointly evaluates document-level prediction and alignment with span-level psycholinguistic evidence (Samory et al., 2026); accordingly, our contribution is a reproducible baseline suite for this shared-task setting rather than a new model family.

Parallel to these developments, research on harmful content detection has increasingly emphasized interpretability. While transformer-based models achieve strong performance on misinformation and propaganda detection (Liu et al., 2023; Ermurachi and Gîfu, 2020), they often lack transparency regarding the linguistic evidence underlying their predictions. Prior work on propaganda detection has shown that explicit linguistic markers can improve interpretability and support human-aligned explanations (Ermurachi and Gîfu, 2020). This perspective resonates with psycholinguistic approaches that categorize discourse elements into interpretable functional roles such as actors, actions, threats, or evidence.

SemEval-2026 Task 10 (PsyCoMark) (Samory et al., 2026) builds on these insights by providing span-level psycholinguistic marker annotations aligned with document-level conspiracy labels.

This design enables the study of how conspiratorial reasoning is linguistically constructed and how interpretable markers can be integrated into computational models. The task is conceptually inspired by prior work that combines linguistic structure with predictive modeling, including semiotic analyses of political discourse (Gîfu and Cristea, 2013) and marker-based approaches to persuasive language. Relative to prior conspiracy-detection benchmarks, our contribution is intentionally that of a reproducible baseline suite for a new shared task that jointly evaluates document prediction and evidence alignment. Taken together, the literature suggests that conspiracy detection requires models capable of capturing both surface-level linguistic cues and deeper psycholinguistic structures. PsyCoMark offers a unique opportunity to evaluate such models in a controlled setting, bridging the gap between interpretability and predictive performance.

3 Dataset and Methods

3.1 Dataset

We use the official PsyCoMark dataset released for SemEval-2026 Task 10, which provides document-level conspiracy labels (Yes/No/Can't tell) and span-level psycholinguistic marker annotations across six interpretable categories: Actor, Action, Victim, Evidence, Threat, and Effect (Samory et al., 2026). Following the task protocol, we rehydrate all instances using the provided identifiers, resulting in three splits: `train_rehydrated.jsonl`, `dev_rehydrated.jsonl`, and `test_rehydrated.jsonl`.

For the binary classification setting, we retain only Yes/No instances, excluding *Can't tell* from training. For the zero-shot evaluation, all three labels are preserved. Minimal preprocessing is applied to maintain linguistic fidelity: URL masking, whitespace normalization, and sentence segmentation using spaCy (Honnibal et al., 2020). No additional filtering or text cleaning is performed, as conspiratorial cues often rely on subtle lexical and pragmatic signals.

To support interpretability analyses, we also extract candidate evidence sentences using lightweight heuristics (modal verbs, interrogatives, generic agentive constructions). These sentences are not used for training but serve as auxiliary material for qualitative inspection.

3.2 Methods

Our approach combines marker-informed feature engineering, supervised transformer baselines, and zero-shot prompting, forming a transparent and reproducible pipeline aligned with the goals of PsyCoMark.

3.2.1 Marker-derived lexicon and LIWC-style features

Using the gold span annotations, we construct a marker lexicon by aggregating token–category pairs and assigning each token its dominant marker type. From this lexicon, we compute LIWC-style features: raw counts and normalized ratios for each marker category. These features provide interpretable psycholinguistic signals and serve as a lightweight baseline for conspiracy detection. For reproducibility and inspection, we release the full lexicon (`psy_dict.csv`) in the repository.

3.2.2 Supervised transformer baselines

We fine-tune RoBERTa and DeBERTa-family encoders for binary Yes/No classification. Small and medium models undergo Optuna-based hyperparameter search (Akiba et al., 2019), exploring learning rate, batch size, warmup ratio, weight decay, and sequence length. Large models (e.g., `microsoft/deberta-v3-large`) are trained under step-bounded regimes with mixed precision and gradient checkpointing to ensure computational feasibility.

All models use cross-entropy loss, linear warmup, and early stopping based on development-set weighted F1.

3.2.3 Heuristic relabeling (ablation)

We experiment with a logistic regression classifier trained on marker-derived features using only confident Yes/No instances. The classifier is then used to relabel *Can't tell* cases under conservative probability thresholds. This heuristic does not improve downstream performance and is excluded from the final system.

3.2.4 Zero-shot TinyLlama baseline

We evaluate `TinyLlama-1.1B-Chat-v1.0` (Team, 2024) in a zero-shot three-way classification setting. Prompts include simple, definition-augmented, and step-by-step variants. Greedy decoding is used for reproducibility. This baseline highlights prompt sensitivity and label bias, consistent with prior findings on compact instruction-tuned models (Zhuo et al., 2024; Renze, 2024).

3.3 System architecture

To support student implementation, we adopt a modular architecture that mirrors standard ACL shared-task pipelines (Fig. 1).

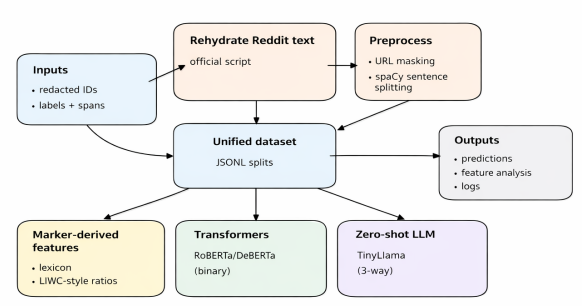

Figure 1: System architecture for the PsyCoMark baseline pipeline. The input consists of rehydrated Reddit posts and span-level psycholinguistic marker annotations from the shared-task dataset.

- **Preprocessing module:** This module handles text acquisition and normalization. It performs rehydration of the PsyCoMark instances, URL masking, whitespace normalization, and sentence segmentation. The goal is to preserve linguistic cues while ensuring consistent input formatting across models.

Listing 1: Preprocessing: candidate sentence extraction (heuristic).

```
import spacy
nlp = spacy.load("en_core_web_sm")

MODALS = {"must","should","could","might","may","would
      ","will","can"}
AGENTS = {"elites","government","they","globalists","
      cabal"}

def extract_candidates(text):
   out = []
   for i, sent in enumerate(nlp(text).sents):
      s = sent.text.strip()
      has_modal = any(t.lemma_.lower() in MODALS for t
            in sent)
      if has_modal or s.endswith("?") or any(t.text.
            lower() in AGENTS for t in sent):
         out.append({"i": i, "segment": s})
   return out
```

- **Marker processing module:** Using the gold span annotations, we construct a token-level marker lexicon and derive LIWC-style psycholinguistic features. These include raw counts and normalized ratios for each marker category (Actor, Action, Victim, Evidence, Threat, Effect). This module provides interpretable, low-dimensional features aligned with the task's annotation schema.

Listing 2: Marker features: LIWC-style counts/ratios from a token→category lexicon.

```
import re
TOKEN_RE = re.compile(r"[A-Za-z']+")

def liwc_features(text, lexicon, categories):
   toks = [t.lower() for t in TOKEN_RE.findall(text)]
   total = max(1, len(toks))
   counts = {c: 0 for c in categories}
   for tok in toks:
      if tok in lexicon:
         counts[lexicon[tok]] += 1
   return {f"{c}_ratio": counts[c]/total for c in
         categories}
```

- **Feature-based classifier:** A logistic regression classifier is trained on the marker-derived features. This shallow model serves as a transparent baseline and enables ablation studies on the predictive value of psycholinguistic markers alone.

Listing 3: Baseline: logistic regression over marker-derived features.

```
from sklearn.linear_model import LogisticRegression

clf = LogisticRegression(max_iter=2000)
clf.fit(X_train, y_train)
pred = clf.predict(X_dev)
```

- **Transformer-based classifier:** We fine-tune RoBERTa and DeBERTa models for binary Yes/No classification. Hyperparameters are optimized via Optuna for small/medium models, while large models use step-bounded training with mixed precision and gradient checkpointing. This module captures contextual semantics beyond marker-level cues.

Listing 4: Transformer fine-tuning (binary) with HuggingFace Trainer.

```
from transformers import AutoTokenizer,
      AutoModelForSequenceClassification, Trainer,
      TrainingArguments

tok = AutoTokenizer.from_pretrained("bert-base-uncased
      ")
model = AutoModelForSequenceClassification.
      from_pretrained("bert-base-uncased", num_labels
      =2)

args = TrainingArguments(output_dir="runs",
      num_train_epochs=3, learning_rate=2e-5,
                  per_device_train_batch_size=16,
                        evaluation_strategy="epoch")

trainer = Trainer(model=model, args=args,
      train_dataset=ds_train, eval_dataset=ds_dev)
trainer.train()
```

- **Zero-shot inference module:** `TinyLlama-1.1B-Chat-v1.0` is used for three-way classification (Yes/No/Can't tell) in a zero-shot setting. Prompts include definition-augmented and step-by-step variants. This module highlights prompt sensitivity and the limitations of compact instruction-tuned models.

Listing 5: Zero-shot TinyLlama prompting (greedy) and label parsing.

```
PROMPT = "Answer with Yes/No/Cant.\nText:\n{txt}\
      nAnswer:"

def predict_label(txt):
   out = lm.generate(**tok(PROMPT.format(txt=txt),
         return_tensors="pt").to(lm.device),
               do_sample=False, max_new_tokens=8)
   gen = tok.decode(out[0], skip_special_tokens=True).
         lower()
```

```
    return "Yes" if "yes" in gen else ("No" if "no" in
        gen else "Cant")
```

4 Results

We evaluate all models on the official PsyCoMark development and test sets. Supervised models are trained in the binary setting (Yes/No), while the zero-shot baseline is evaluated on the full three-way classification task. All transformer models are selected based on development-set weighted F1. For the supervised baselines, we report the best verified run per model on the development split and for the blind test split, we report the official score returned for the submitted system.

4.1 Dev Results (Best per Model)

Table 1 summarizes the best development weighted F1 achieved by each evaluated transformer model. The best dev model is `microsoft/deberta-v3-large`.

Table 1: Best weighted F1 on the development dataset (from the baseline write-up). Large models use a time-bounded, step-based training configuration.

Model	Alias	Dev Weighted F1
DistilRoBERTa	distilroberta	0.7614
DeBERTa-v3 Small	deberta_small	0.7581
RoBERTa Base	roberta_base	0.7465
DeBERTa-v3 Base	deberta_base	0.7496
RoBERTa Large	roberta_large	0.7640
DeBERTa-v3 Large	deberta_large	**0.8339**

4.2 Best Hyperparameter Configurations

Table 2 and Table 3 report the best configurations for small/medium models found with Optuna, and Table 4 reports fixed, time-bounded configurations used for large models. Abbreviations: L=max length, LR=learning rate, WD=weight decay, WARM=warmup ratio, BS=per-device batch, GA=gradient accumulation, EP=epochs, GC=gradient checkpointing.

Table 2: Optuna configs (part A): sequence + optimizer.

Alias	L	lr	wd	warm
distilroberta	320	5.88e-06	0.0366	0.1075
deberta_small	256	4.30e-06	0.0156	0.0399
roberta_base	320	5.88e-06	0.0366	0.1075
deberta_base	256	4.30e-06	0.0156	0.0399

Table 3: Optuna configs (part B): batch/epochs/checkpointing.

Alias	bs	ga	ep	gc
distilroberta	8	2	6	False
deberta_small	8	1	3	True
roberta_base	8	2	6	False
deberta_base	8	1	3	True

Table 4: Fixed large-model configs (step-bounded).

Alias	L	steps	lr	wd	warm	bs	ga	ep	gc
roberta_large	192	400	1e-05	0.01	0.06	2	16	1	True
deberta_large	192	400	1e-05	0.01	0.06	2	16	1	True

4.3 Main Results

Cross-model comparison is reported on the development split (Table 1). For the blind evaluation split, the official score available to us is the submitted system `microsoft/deberta-v3-large` with Test Weighted F1 0.75.

The marker-only model confirms that psycholinguistic cues alone are insufficient for high-accuracy detection. Transformer models capture richer contextual semantics, with large-capacity models consistently outperforming base variants on the development split. The improvement from RoBERTa-base to RoBERTa-large and from DeBERTa-v3-base to DeBERTa-v3-large indicates that model capacity plays a meaningful role in capturing subtle conspiratorial cues.

4.4 Ablation: Marker-Driven Relabeling

We also tested heuristic relabeling of *Can't tell* instances using a logistic regression classifier trained on marker-derived features. The relabeling variant obtained Dev Weighted F1 0.7605. The relabeling variant did not improve over the no-relabeling baseline, so it was not used in the final submission.

4.5 Zero-Shot Prompting Results

As a compact sanity-check baseline, we verified a reproducible TinyLlama zero-shot setting using a definition-augmented prompt and greedy decoding on a 500-example sample. This setup reaches 0.34 accuracy, confirming substantial majority-class bias and weak handling of *Can't tell* cases in the three-way setting.

Zero-shot performance is low, with strong majority-class bias and inconsistent handling of

Can't tell cases showing that compact instruction-tuned models struggle with fine-grained moderation tasks requiring nuanced contextual reasoning (Zhuo et al., 2024; Renze, 2024).

4.6 Error Analysis

Across models, false positives often involve emotionally charged but non-conspiratorial content, while false negatives correspond to implicit conspiratorial reasoning lacking explicit markers. This pattern reinforces the need for models that integrate both contextual semantics and interpretable psycholinguistic cues.

5 Conclusion

This paper addressed the central question of how interpretable psycholinguistic markers can be integrated into a conspiracy-detection system that is accurate, reproducible, and transparent. Our baselines show that marker-derived features provide clear, human-aligned evidence signals but remain insufficient on their own, while transformer-based models—especially DeBERTa-v3-large—capture the deeper contextual cues required for strong predictive performance.

By combining marker-aware representations with high-capacity encoders, the proposed pipeline demonstrates that interpretability and accuracy can be jointly pursued within the PsyCoMark framework. These baselines establish a controlled reference point for the shared task and highlight the need for future systems that explicitly integrate marker extraction with supervised three-way classification, advancing psycholinguistically grounded and reliable moderation workflows.

Acknowledgments

This work was carried out partially within the project "Tools for Processing Online Texts Specific to Cultural and Scientific Diplomacy", funded by the Academy of Romanian Scientists.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Jovan Byford. 2011. *Conspiracy Theories: A Critical Introduction*. Palgrave Macmillan.

Francesco Corso, Francesco Pierri, and Gianmarco De Francisci Morales. 2025. Conspiracy theories and where to find them on tiktok. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8346–8362.

Karen M. Douglas, Robbie M. Sutton, and Aleksandra Cichocka. 2017. The psychology of conspiracy theories. *Current Directions in Psychological Science*, 26(6):538–542.

V. Ermurachi and Daniela Gîfu. 2020. Uaic1860 at semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of SemEval*, pages 1835–1840.

Amos Fong, Jon Roozenbeek, Danielle Goldwert, Steven Rathje, and Sander van der Linden. 2021. The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on twitter. *Group Processes & Intergroup Relations*, 24(4):606–623.

Daniela Gîfu. 2013. *Temeliile Turnului Babel: O perspectivă integratoare asupra discursului politic*. Editura Academiei Române.

Daniela Gîfu. 2016. *Lexical Semantics in Text Processing: Contrastive Diachronic Studies on Romanian Language*. Placeholder reference (publisher/details to be completed).

Daniela Gîfu and Dan Cristea. 2012. Multidimensional analysis of political language. In *Lecture Notes in Electrical Engineering*, volume 164, pages 213–221. Springer.

Daniela Gîfu and Dan Cristea. 2013. Towards an automated semiotic analysis of the romanian political discourse. *Computer Science Journal of Moldova*, 21(1):36–64.

Matthew Honnibal and 1 others. 2020. spacy: Industrial-strength natural language processing in python. Software documentation.

Hui Liu, Wenya Wang, and Haoliang Li. 2023. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9781–9796, Toronto, Canada. Association for Computational Linguistics.

Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2022. Loco: The 88-million-word language of conspiracy corpus. *Behavior Research Methods*, 54:1794–1817.

Samantha C. Phillips, Lynnette Hui Xian Ng, and Kathleen M. Carley. 2022. Hoaxes and hidden agendas: A twitter conspiracy theory dataset: Data paper. In *Companion Proceedings of the Web Conference 2022*, Virtual Event, Lyon, France. ACM.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.

Mattia Samory, Felix Soldner, and Veronika Batzdorfer. 2026. SemEval-2026 task 10: PsyCoMark – psycholinguistic conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, 34(3):150–160.

TinyLlama Team. 2024. Tinyllama-1.1b-chat-v1.0. Model card / release information.

Ştefan Vlăduţescu, Florentin Smarandache, Daniela Gîfu, and Alexandru Ţenescu. 2014. *Topical Communication Uncertainties*. Sitech & Zip Publishing.

Jing Zeng and D. Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on tiktok. *Policy & Internet*, 14(1):79–95.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. Prosa: Assessing and understanding the prompt sensitivity of llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.