OPINE: A Prior-calibrated Scoring Framework for LLM-based Multi-label Scientific Opinion Classification

Mengting Zhang[1,2], Gaofeng Pan[3], Zhixiong Zhang[1,2] *, Yang Li[1,2], Guangyin Zhang [1,2]
[1]National Science Library, Chinese Academy of Sciences
[2] Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Science
[3] School of Information Management, Zhengzhou University
zhangmengting@mail.las.ac.cn, pangaofeng@stu.zzu.edu.cn, zhangzhx@mail.las.ac.cn, liyang2022@mail.las.ac.cn, zhangguangyin@mail.las.ac.cn

Abstract

Scientific opinion classification based on discourse functions provides a structured semantic basis for analytical tasks such as gap identification and hypothesis generation. However, this task is uniquely challenged by the multi-label nature of scientific expressions and AIMRaD structural constraints. Existing LLM-based methods typically rely on direct label generation, which obscures decision logic, or treat discourse information as passive context rather than a structural prior. We propose **OPINE**, a multi-stage framework that reformulates classification as a controllable *scoring-calibration-refinement* pipeline. By decoupling textual evidence from decision logic, OPINE generates independent label-wise affinity scores calibrated by AIMRaD priors. To resolve the multi-label challenge, we introduce a quantile-based decoding rule to naturally capture co-existing roles, alongside a pairwise refinement mechanism to mitigate confusion between similar categories. We contribute a new benchmark of 18 discourse functions across diverse sections. Experimental results show that OPINE generally outperforms strong baselines, reaching F1 scores of 63.20%, 53.68%, and 63.22% under Micro, Macro, and Example settings, respectively. Our analysis reveals that integrating discourse structures as explicit priors is superior to conventional passive context integration, while pairwise refinement successfully mitigates confusion between functionally similar categories. The code and dataset are available at https://github.com/znoodle63/OPINE.

1 Introduction

Scientific argumentation involves more than the objective reporting of empirical facts (Kando, 1997): authors introduce evidence-grounded subjective judgments about research problems, methods, and results (Ni et al., 2024; Zhang et al., 2025b), which transform raw observations into knowledge claims.

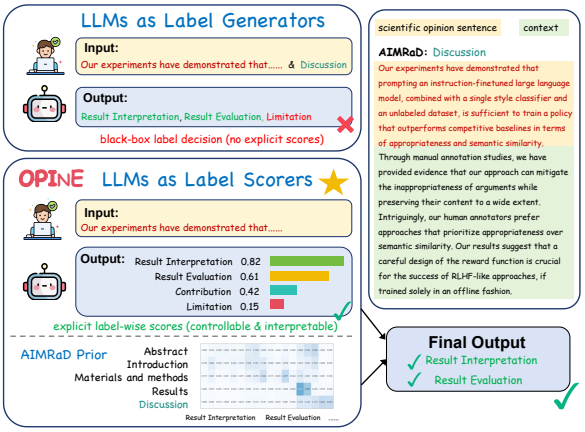

Figure 1: Comparison between the *LLMs as Label Generators* paradigm and our proposed *LLMs as Label Scorers* framework. OPINE replaces a single black-box label decision with explicit label-wise scores and AIMRaD-guided inference.

We term these ***scientific opinions***, which manifest in statements such as those evaluating work, interpreting findings, or reflecting on limitations (Teufel, 1999). Rather than mere linguistic variants, these opinions instantiate specific discourse functions that structure scientific reasoning (Teufel et al., 1999). Formally classifying them clarifies their reasoning roles, providing a semantic basis for analytical tasks including gap identification (Rabiei et al., 2017), hypothesis generation (Curtis, 2012; Alkan et al., 2025), and future research agenda setting (Jagannathan et al., 2023).

Despite its significance, classifying scientific opinions by discourse functions remains underexplored. Prior work predominantly targets user-centric contexts (e.g., social media, product reviews) (Wankhade et al., 2022), focusing on sentiment polarity (Safira et al., 2025) or stance (Yang et al., 2025) through sentence-level single-label classification. However, scientific opinions defy this formulation due to two core properties. First, they are functionally composite: a single sen-

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 32313–32333
July 2-7, 2026

tence often performs multiple functions concurrently (Zhang et al., 2025b), such as interpreting results while evaluating methods. Second, their distribution is discourse-conditioned: the AIMRaD (***A**bstract, **I**ntroduction, **M**aterials and methods, **R**esults, **a**nd **D**iscussion*) structure (Teufel, 2010; Cargill and O'Connor, 2013) leads to section-dependent concentrations and long-tailed label distributions. Consequently, we formulate ***scientific opinion classification*** as a discourse-aware multi-label task, where label behavior is constrained by broader discourse context rather than isolated sentence-level features.

To address these task requirements, Large Language Models (LLMs) are a natural choice, given their strong generalization in low-resource and imbalanced settings (Zhao et al., 2023; Kojima et al., 2022; Hu et al., 2025). However, existing LLM-based approaches face several limitations. Most prompt-based methods follow an ***LLMs as Label Generators*** paradigm (Zhu and Zamani, 2024; Yoshimura and Kashima, 2025; Hu et al., 2025), which collapses the label-wise decision process into a single black-box generation step, concealing the model's relative confidence and the rationale behind label selection. Furthermore, AIMRaD discourse information is typically treated as passive context rather than a structural prior, failing to directly guide the decision-level logic. Finally, LLMs often struggle to distinguish categories that are semantically similar but functionally distinct (Zhu et al., 2024), leading to confusion between fine-grained labels and revealing room for improvement in nuanced category discrimination.

To bridge these gaps, we propose **OPINE** (**O**pinion **P**rior-aware **I**nference with **N**uanced **E**nhancement), a multi-stage framework that reformulates scientific opinion classification as a controllable *scoring-calibration-refinement* pipeline. Departing from the conventional *LLMs as Label Generators* paradigm, OPINE formalizes an ***LLMs as Label Scorers*** approach that decouples textual evidence from decision logic. It first produces explicit label-wise affinity scores, which are then probabilistically calibrated by integrating AIMRaD-based structural priors at the decision level. To reconcile the discrepancy between continuous scores and discrete categories, we introduce a data-driven quantile-based decoding rule to naturally capture multiple co-existing discourse functions, and a pairwise refinement mechanism to sharpen decision boundaries and resolve marginal confusion between functionally similar candidates.

Our main contributions are as follows:

- We propose **OPINE**, a novel multi-stage framework that reformulates scientific opinion classification as a decoupled *scoring-calibration-refinement* pipeline to address the limitations of direct LLM-based label generation.
- We construct a new evaluation benchmark that categorizes scientific opinions into 18 classes based on *discourse functions*, while providing rich contextual metadata and AIMRaD-based section information for each instance.
- We show that OPINE generally outperforms strong baselines, demonstrating that explicit discourse priors are superior to conventional context integration and that pairwise refinement effectively mitigates marginal confusion between similar categories.

2 Related Works

2.1 LLM-Based Generative Approaches to Multi-Label Classification

Large Language Models (LLMs) have demonstrated strong generalization capabilities for multi-label text classification, particularly in low-resource and few-shot scenarios where annotated data is scarce (Hu et al., 2025; Abdeen et al., 2025). To adapt LLMs for this task, existing research predominantly employs generative prompting strategies. Xia et al. (2025) categorize these into three main types: *flattened* generation, which selects from the full label list; *per-parent*, a top-down recursive approach; and *per-path*, which generates complete hierarchical paths. Beyond structural formatting, recent studies have enhanced performance through confidence-ranked reasoning (Yu and Wang, 2025), semantic label augmentation (Zhang et al., 2025c), and ensemble frameworks (Zhu and Zamani, 2024; Sakai and Lam, 2025). Despite these advancements, Ma et al. (2025) note that autoregressive generation tends to suppress concurrent labels, often obscuring the model's true underlying confidence. This limitation highlights the potential of transitioning from direct text generation toward more explicit scoring-based inference.

2.2 LLM-as-Scorer Methods in Ranking and Evaluation Tasks

Beyond standard text generation, the LLM-as-a-Scorer paradigm has recently been explored in

Opinion Object	Category
A. Scientific Community	
A.1 Research Actors	A.1-1 Opinions on the contribution and influence of research actors
A.2 Research Outputs	A.2-1 Opinions on the value and impact of research outputs
A.3 Research Objects	A.3-1 Opinions on the overall status and development stage of a research field
	A.3-2 Opinions on current key research priorities and hotspots
	A.3-3 Opinions on mainstream knowledge and consensus regarding a research topic
	A.3-4 Opinions on mainstream technical routes in the field
	A.3-5 Opinions on current research gaps and insufficiencies in existing research
	A.3-6 Opinions on current research difficulties and challenges faced by current research
	A.3-7 Opinions on the selection and formulation of research problems
	A.3-8 Opinions on the value of a research field, topic, or specific problem
B. Authors' Own Research	
B.1 Research Hypotheses	B.1-1 Opinions concerning research hypotheses
B.2 Research Methods	B.2-1 Opinions on the soundness and rationale of methodological design
B.3 Experimental Design	B.3-1 Opinions on the soundness and rationale of experimental design
B.4 Research Results	B.4-1 Opinions that interpret research results or provide causal explanations
	B.4-2 Opinions that compare research results and judge superiority or effectiveness
B.5 Research Conclusions	B.5-1 Opinions on the contributions and value of the authors' own research
	B.5-2 Opinions on the limitations and insufficiencies of the authors' own research
	B.5-3 Opinions on future improvements and development directions

Table 1: Taxonomy of scientific opinions.

several non-classification contexts to support finer-grained inference. In Information Retrieval, LLMs have been used as zero-shot rankers, where pointwise, pairwise, and setwise scoring has been applied to estimate relevance and improve relative ranking (Zhuang et al., 2023; Li et al., 2025). In Natural Language Evaluation, probabilistic scoring has been used to assess dialogue quality and student assignments, and has been shown to offer greater stability than direct text generation (Chen et al., 2024; Chiang et al., 2024), with further benefits for calibration and uncertainty estimation (Sun et al., 2025). However, existing applications of scoring-based inference remain largely confined to ranking and evaluation tasks, and its potential within scientific opinion classification remains underexplored.

3 Taxonomy of Scientific Opinions

Scientific opinions refer to evidence-grounded subjective judgments that authors make about specific scientific objects (Zhang et al., 2025a,b). Building on this definition, we classify scientific opinions by jointly considering the object they concern and the discourse functions they fulfill in scientific reasoning. These functions include value judgment, limitation analysis, interpretation or causal explanation, and forward-looking inference. Following this principle, we construct a taxonomy that distinguishes opinions directed toward the **scientific community** (Group A) from those concerning the **authors' own research** (Group B). Within each group, scientific opinions are further organized by the scientific objects involved, including research fields, topics, problems, methods, experimental design, and results, yielding 18 well-defined opinion types as shown in Table 1. Detailed definitions of each type are provided in Appendix B.3.

4 Methodology

4.1 Overview

We propose **OPINE** (**O**pinion **P**rior-aware **I**nference with **N**uanced **E**nhancement), a multi-stage framework that reformulates scientific opinion classification as a controllable *scoring-calibration-refinement* pipeline. Initially, the system decouples label evidence from the final decision by prompting LLMs to generate independent, text-based affinity scores P_{text} for each candidate label (§4.2). These scores are subsequently calibrated by integrating the AIMRaD structure as a statistical prior P_{prior}, yielding a discourse-aware posterior distribution P^* (§4.3). To determine the final multi-label set, we employ a quantile-based decoding rule that identifies a sparse set of active labels while preserving a stable primary prediction (§4.4). Finally, to resolve the marginal confusion often remaining between top-ranking candidates, a pairwise refinement mechanism acts as a local auditor to shrink the decision margin for pre-defined

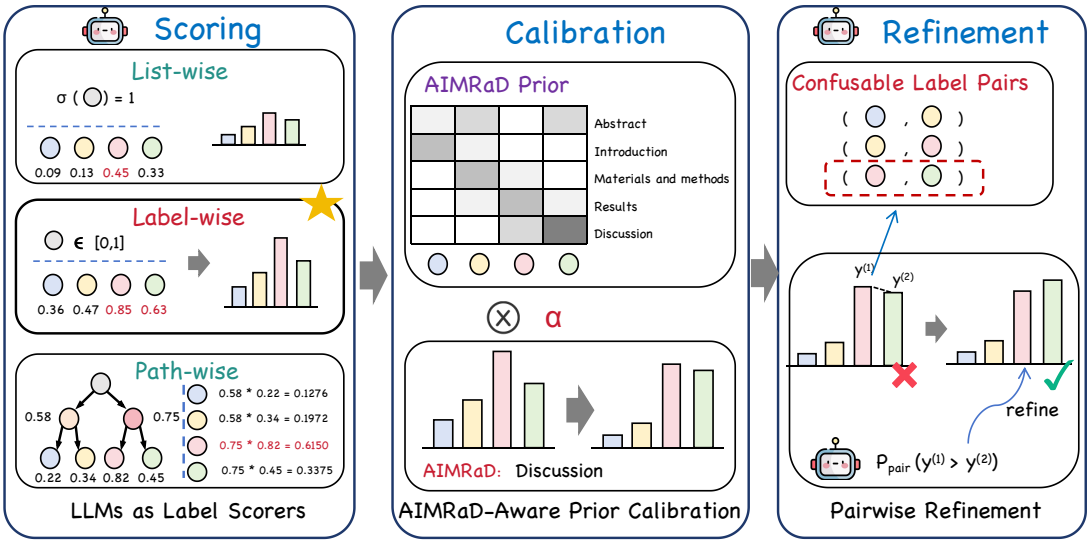

Figure 2: Overall architecture of our proposed OPINE framework. The model operates in three stages: (1) label-wise scoring with LLMs, (2) AIMRaD-aware prior calibration, and (3) pairwise refinement for confusable labels .

confusable pairs, ensuring nuanced category discrimination (§4.5).

4.2 LLMs as Label Scorers

Rather than prompting LLMs to generate a label set directly, we reformulate the task as a label-wise scoring problem. Given an input sentence x and its local context c, we construct a prompt that pairs (x, c) with each candidate label $y \in \mathcal{Y}$, and LLMs are prompted to estimate the extent to which the sentence expresses the semantic function associated with y. The LLMs' responses are subsequently mapped to scalar values, producing text-based affinity scores $P_{\text{text}}(y \mid x, c)$. These scores are not treated as posterior probabilities of a predictive classifier; instead, they indicate how strongly the sentence instantiates the function encoded by y, based solely on textual evidence.

This scoring-based formulation makes the decision process explicit by decoupling *label scoring* from *label selection*, exposing comparable cross-label evidence and yielding a controllable intermediate representation.

4.3 AIMRaD-Aware Prior Calibration

Scientific papers are typically organized according to established discourse structures, among which the AIMRaD structure is the most widely adopted organizational framework in contemporary scientific writing (Cargill and O'Connor, 2013). This structure partitions the research narrative into sections with clearly defined functional roles: *Introduction* provides the background and introduces the research problem, *Materials and Methods* describes the study design and implementation process, *Results* reports empirical findings and associated evidence, and *Discussion* interprets the results and reflects on their implications and limitations. Under this functional structure, different types of scientific opinions tend to exhibit section-dependent distribution patterns, rather than occurring randomly or uniformly across the paper.

Motivated by this observation, we treat AIMRaD as a structural prior that complements and calibrates the text-based affinity scores, thus imposing discourse-level constraints. Concretely, we estimate a section-conditioned prior probability from the empirical distribution of each fine-grained label across sections in the annotated corpus.

$$P_{\text{prior}}(y \mid s) = \frac{\text{count}(y, s)}{\sum_{y' \in \mathcal{Y}} \text{count}(y', s)} \quad (1)$$

Here, s denotes the AIMRaD section to which the sentence belongs, and $P_{\text{prior}}(y \mid s)$ characterizes the structural tendency of label y to occur in that section. This prior acts as a *soft structural constraint* that calibrates the decision process.

To calculate the posterior distribution $P^*(y \mid x, c, s)$, we assume a mild conditional independence assumption (see Appendix C), under which the posterior is proportional to the text-based affinity scores $P_{\text{text}}(y \mid x, c)$ and the structural prior $P_{\text{prior}}(y \mid s)$:

$$P^*(y \mid x, c, s) \propto P_{\text{text}}(y \mid x, c)\, P_{\text{prior}}(y \mid s) \quad (2)$$

where s is the AIMRaD section. However, the relative importance of these two sources of information may vary across datasets and application scenarios. To account for this, we introduce a tunable balancing parameter $\alpha \in (0, 1]$, and obtain the final posterior by weighting the two components and normalizing over the label space:

$$P^*(y \mid x, c, s) = \frac{P_{\text{text}}(y \mid x, c)^{\alpha} P_{\text{prior}}(y \mid s)^{1-\alpha}}{\sum_{y' \in \mathcal{Y}} P_{\text{text}}(y' \mid x, c)^{\alpha} P_{\text{prior}}(y' \mid s)^{1-\alpha}} \quad (3)$$

where α controls the trade-off between textual evidence and structural priors: $\alpha \to 1$ emphasizes text-based evidence, whereas $\alpha \to 0$ drives predictions toward the AIMRaD prior.

4.4 Quantile-Based Label Decoding

After obtaining the discourse-aware posterior distribution $P^*(y \mid x, c, s)$, we convert continuous probabilities into a discrete label set using a quantile-based decoding rule, rather than a fixed threshold or top-k selection. We first retain the top-1 label

$$y^{(1)} = \arg\max_{y \in \mathcal{Y}} P^*(y \mid x, c, s) \quad (4)$$

ensuring that every instance receives at least one prediction. We then collect the probabilities of all non-top-1 labels over the dataset to form a background distribution $\mathcal{B}$, and compute a global threshold $\tau = \text{Quantile}_q(\mathcal{B})$, where q controls the selectivity of additional labels. The final prediction set is defined as

$$\hat{Y} = \{y^{(1)}\} \cup \{\, y \neq y^{(1)} \mid P^*(y \mid x, c, s) \geq \tau \,\} \quad (5)$$

This decoding strategy preserves a stable primary decision while adding extra labels only when their posterior evidence exceeds an empirically estimated threshold, avoiding manual tuning and making the decision process more interpretable.

4.5 Pairwise Refinement for Confusable Labels

To further mitigate the marginal confusion between semantically similar yet functionally distinct categories, we introduce a Pairwise Refinement mechanism. The choice to focus on the top-2 candidates is inherently tied to our quantile-based decoding (Eq.5): since the strategy typically yields a sparse set of active labels, the most critical decision boundary often lies between the primary candidate $y^{(1)}$ and the most competitive runner-up $y^{(2)}$.

We define a set of confusable pairs $\mathcal{C} = \{(y_i, y_j) \mid y_i, y_j \in \mathcal{Y}\}$, representing labels that are historically difficult for the model to disambiguate. For any instance, the refinement layer is activated only if the predicted top-2 pair $\{y^{(1)}, y^{(2)}\}$ exists in $\mathcal{C}$. Instead of a simple re-ranking, we implement a *Margin Shrinkage* strategy to calibrate the decision. Concretely, we prompt the LLM to perform a pairwise comparison, yielding a preference probability $P_{\text{pair}}(y^{(1)} \succ y^{(2)})$.

To ensure stable and interpretable refinement, we follow two design principles: (i) *neutral consistency*, where $P_{\text{pair}} = 0.5$ leaves the original margin unchanged, and (ii) *probability mass preservation*, which keeps the total score of the top pair invariant. Let $S = P^*(y^{(1)}) + P^*(y^{(2)})$ denote the total score and $m = P^*(y^{(1)}) - P^*(y^{(2)})$ the initial margin. The margin is adjusted as:

$$m' = 2 \cdot P_{\text{pair}}(y^{(1)} \succ y^{(2)}) \cdot m \quad (6)$$

The refined scores s'_1 and s'_2 are then reconstructed:

$$s'_1 = \frac{S + m'}{2}, \quad s'_2 = \frac{S - m'}{2} \quad (7)$$

By shrinking m towards zero, this mechanism effectively softens the primary model's overconfidence. In our quantile-based framework, this increase in s'_2 allows the runner-up to more easily exceed the threshold τ, facilitating a more nuanced multi-label prediction.

5 Experiments

5.1 Experimental setup

Datasets and Evaluation Metrics. We construct a new benchmark for fine-grained scientific opinion classification, comprising 5,024 opinion sentences annotated from 82 full-text papers in ACL 2024 [1]. Following our two-level taxonomy (§3; details in Appendix D), each instance is categorized into 18 classes based on target objects and discourse functions, and enriched with contextual metadata and AIMRaD section information. The dataset was annotated by two PhD students, achieving a Cohen's Kappa of 0.863, indicating a high level of inter-annotator agreement. To ensure a fair and leakage-free evaluation, we adopt a strict paper-level splitting strategy, partitioning the dataset into 66 training papers and 16 test papers (approximately 4:1), such that sentences from the same

[1] https://aclanthology.org/events/acl-2024/

Model	Micro			Macro			Example		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Phi-4	48.13	53.87	50.84	41.04	37.50	35.01	48.81	54.66	50.40
Ministral-3-14B-Instruct-2512	45.85	51.32	48.43	38.84	36.95	34.21	48.05	51.89	48.89
Qwen-3-14B	51.30	57.43	54.19	53.56	45.96	43.80	53.38	57.96	54.43
Qwen-3-32B	**55.54**	**62.17**	**58.67**	**55.65**	**48.54**	**45.85**	**57.58**	**62.90**	**58.91**
DeepSeek-V3.2<deepseek-chat>	55.37	61.99	58.49	52.97	48.91	46.29	56.43	62.49	58.13
DeepSeek-V3.2<deepseek-reasoner>	55.62	62.26	58.75	53.73	47.11	45.03	56.96	62.92	58.44
GPT-5	58.96	66.00	62.28	57.05	53.57	51.72	60.49	66.71	62.06
Gemini-2.5-Pro	**59.80**	**67.00**	**63.20**	**61.30**	**56.00**	**53.68**	**61.73**	**67.76**	**63.22**

Table 2: Final results of all models under our proposed **OPINE**. We report Micro-, Macro-, and Example-average precision, recall, and F1. The overall best result is **bold** and the second best is underlined. For open-weight models, the best result is **bold**, and the second best is underlined.

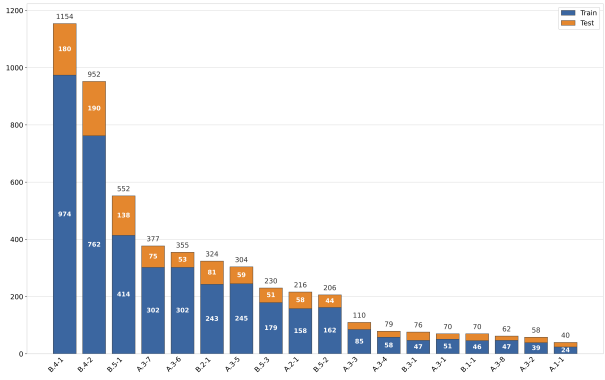

Figure 3: Distribution of Scientific Opinion Types

paper do not appear in both splits. All 18 labels are covered in both the training and test sets, and the label distribution across splits is illustrated in Figure 3. Following previous works (Zhang et al., 2025d; Xia et al., 2025), we employ Micro-F1, Macro-F1, and Example-F1 as metrics to evaluate overall performance.

Baselines. To evaluate the effectiveness of OPINE, we compare it against the conventional paradigm and analyze internal variations within our proposed framework. (1) **LLMs as Label Generators**. We implement three common strategies based on the ensembling and prompting framework of Xia et al. (2025) to serve as our primary baselines, where models are tasked with directly generating category names: **Flattened**. LLMs generate labels directly from a flat list of 18 categories. **Per-parent**. A top-down approach where the model first selects a parent group (Group A or B) before determining the specific category. **Per-path**. Models are prompted to generate the complete hierarchical path for each identified opinion. (2) **LLMs as Label Scorers**. Unlike the generative approach in Xia et al. (2025), we propose a scoring-based inference. We investigate three architectural variations for obtaining affinity scores: **Label-wise**. Scoring each of the 18 labels independently. As this strategy consistently demonstrates the most robust performance across base models (see Table 4), it serves as the backbone of the final OPINE framework. **List-wise**. Scoring all candidate labels within a single prompt to capture inter-label relationships. **Path-wise**. Scoring parent groups before sub-categories to leverage the taxonomy's structural constraints.

We implement these paradigms across eight representative LLMs, including GPT-5 (gpt-5-2025-08-07), Gemini-2.5-Pro, DeepSeek-V3.2 (deepseek-chat and deepseek-reasoner), Qwen-3 (32B and 14B), Phi-4 (14B), and Ministral-3-14B-Instruct-2512. All models are evaluated under three input configurations: **+Context** (incorporating surrounding sentences), **+AIMRaD** (utilizing section metadata), and **+All**. Notably, while the baseline paradigms treat **AIMRaD** information as extra-textual context within the prompt, OPINE uniquely leverages it to construct discourse-aware priors.

Implementation Details. Open-weight models are deployed on NVIDIA A100 (80GB) GPUs, while proprietary models are accessed via official APIs. To ensure consistency across all tasks, we set the decoding temperature to 0.2 for both scoring and generation paradigms. For our prior calibration mechanism, we utilize a quantile threshold of 0.99, and we report results with the prior fusion weight α set to 0.8, which consistently yields optimal performance across most base models. We pre-defined a set of confusable pairs ($\mathcal{C}$) based on functional overlap, including: {(A.3-1, A.3-2), (A.3-1, A.3-3), (A.3-1, A.3-4), (A.3-3, A.3-4), (A.3-5, A.3-6), (A.3-5, A.3-7), (A.3-6, A.3-7), (B.4-1, B.4-2), (B.2-1, B.3-1), (B.5-1, B.5-2), (B.5-2, B.5-3)}.

Paradigm	Strategy	Setting	Micro			Macro			Example		
			P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
LLMs as Label Generators	Flattened		15.36	**80.04**	25.77	14.05	**79.03**	22.55	19.48	80.16	29.68
		+ Context	16.67 ↑	78.67 ↓	27.51↑	15.66 ↑	76.62 ↓	24.29 ↑	20.50 ↑	78.90 ↓	31.02 ↑
		+ AIMRaD	16.73 ↑	79.95 ↓	27.67 ↑	14.87 ↑	77.32 ↓	23.73 ↑	21.55 ↑	**80.27** ↑	32.08 ↑
		+ All	17.96 ↑	79.76↓	29.31 ↑	16.12 ↑	75.92 ↓	25.09 ↑	21.98 ↑	79.81 ↓	32.84 ↑
	Per-parent		45.09	40.20	42.51	41.13	35.28	32.44	33.63	40.20	35.51
		+ Context	45.75 ↑	37.74 ↓	41.36 ↓	35.33 ↓	30.83 ↓	29.06 ↓	31.57 ↓	37.86 ↓	33.44 ↓
		+ AIMRaD	49.84 ↑	42.39 ↑	45.81 ↑	42.46 ↑	35.97 ↑	31.83 ↓	36.05 ↑	42.63 ↑	37.99 ↑
		+ All	49.01 ↑	42.66 ↑	45.61 ↑	38.51 ↓	36.60 ↑	32.52 ↑	36.46 ↑	43.03 ↑	38.32 ↑
	Per-path		23.75	69.92	35.46	19.71	61.61	28.42	28.66	69.73	38.66
		+ Context	25.47 ↑	74.75 ↑	38.00 ↑	22.22 ↑	66.51 ↑	31.38 ↑	30.33 ↑	74.56 ↑	41.02 ↑
		+ AIMRaD	25.85 ↑	75.11 ↑	38.46 ↑	22.48 ↑	65.90 ↑	30.81 ↑	30.33 ↑	74.56 ↑	42.10 ↑
		+ All	26.18 ↑	77.12 ↑	39.09 ↑	22.52 ↑	67.50 ↑	31.81 ↑	30.83 ↑	77.22 ↑	41.93 ↑
LLMs as Label Scorers	Label-wise		50.04	56.15	52.92	52.81	47.77	44.98	52.24	56.47	53.11
		+ Context	51.82 ↑	58.25 ↑	54.85 ↑	54.87 ↑	48.74 ↑	44.88 ↓	53.96 ↑	58.90 ↑	55.10 ↑
	List-wise		50.04	56.06	52.88	46.77	47.44	43.51	52.46	56.32	53.32
		+ Context	51.63 ↑	57.83 ↑	54.56 ↑	46.73 ↓	49.16↑	44.83 ↑	53.86 ↑	58.32 ↑	54.91 ↑
	Path-wise		46.01	51.50	48.60	46.69	44.16	40.70	47.90	52.03	48.87
		+ Context	42.96 ↓	48.13 ↓	45.40 ↓	49.44 ↑	42.41 ↓	39.66 ↓	43.61 ↓	48.60 ↓	44.92 ↓
	OPINE		**55.54**	62.17	**58.67**	**55.65**	48.54	**45.85**	**57.58**	62.90	**58.91**

Table 3: Main results of different LLM-based paradigms under various configurations (Qwen-3-32B). The best result is **bold** and the second best is underlined. ↑means the improvement of over the baseline configuration that uses only sentence content information; while ↓means the decrease. OPINE is built upon the context-enhanced settings .

5.2 Main Results

To evaluate the effectiveness of OPINE, we conduct extensive experiments across eight representative LLMs. The overall performance across all models is summarized in Table 2, and detailed analyses are further provided for four open-weight models, including Qwen-3-32B (Table 3), Qwen-3-14B (Table 6), Phi-4 (Table 7), and Ministral-3-14B-Instruct-2512 (Table 8). Our analysis yields the following key findings:

Performance Gap Between Proprietary and Open-weight Models. Our experiments reveal that proprietary models achieve the strongest overall performance, with Gemini-2.5-Pro ranking first and GPT-5 second. Gemini-2.5-Pro attains 63.20% Micro-F1, 53.68% Macro-F1, and 63.22% Example-F1, whereas the best open-weight model, Qwen-3-32B, achieves 58.67%, 45.85%, and 58.91%, respectively. This gap indicates that proprietary backbones still retain a clear advantage in fine-grained scientific opinion classification, while Qwen-3-32B remains the most competitive open-weight model under the OPINE framework.

Superiority of Scoring over Generation. Our experiments demonstrate that the scoring-based paradigm is a highly effective and stable foundation for scientific opinion classification, generally outperforming generative approaches. As Table 3 shows, Qwen-3-32B achieves a Macro-F1 of 44.98% when acting as a basic *Label-wise scorer*, surpassing most complex generative strategies. This suggests that requesting affinity scores allows the model to capture fine-grained semantic nuances better than discrete text completion, providing a more robust backbone for opinion classification.

Effectiveness of AIMRaD-aware Priors. Our framework, OPINE, demonstrates the significant advantage of integrating AIMRaD-based structural priors into the scoring process, generally leading to substantial performance gains. The most dramatic improvements are observed on Phi-4 (Table 7), where OPINE elevates the Micro-F1 from 42.05% (*Label-wise + Context*) to 50.84%, a remarkable 8.79-point absolute gain. Similarly, for Qwen-3-14B (Table 6), our framework achieves a Micro-F1 of 54.19%, exceeding the best generative baseline (*Per-parent + All*) of 42.99%. While the performance peak varies across architectures, these results indicate that for the majority of tested models, transforming structural metadata into probabilistic priors is far more effective than merely providing it as raw textual context.

Robustness Across Model Scales. By evaluating models of different scales within the same family, such as Qwen-3-14B (Table 6) and Qwen-3-

Strategy	Setting	Qwen-3-32B			Phi-4			DeepSeek-V3.2 <deepseek-reasoner>		
		Micro-F1(%)	Macro-F1(%)	Example-F1(%)	Micro-F1(%)	Macro-F1(%)	Example-F1(%)	Micro-F1(%)	Macro-F1(%)	Example-F1(%)
Label-wise		52.92	44.98	53.11	39.09	33.91	37.72	55.44	47.90	55.67
	+ Context	54.85 ↑	44.88 ↓	55.10 ↑	42.05 ↑	33.89 ↓	41.94 ↑	**59.31** ↑	**51.66**↑	**59.72** ↑
List-wise		52.88	43.51	53.32	36.89	30.84	35.75	41.63	32.01	41.63
	+ Context	54.56 ↑	44.83 ↑	54.91 ↑	36.36 ↓	28.81 ↓	35.07 ↓	38.92 ↓	30.48 ↓	38.75 ↓
Path-wise		48.60	40.70	48.87	33.89	24.62	32.24	44.82	34.35	44.65
	+ Context	45.40 ↓	39.66 ↓	44.92 ↓	33.62 ↓	25.90 ↑	32.35 ↑	46.71 ↑	36.76 ↑	46.47 ↑
OPINE		**58.67**	**45.85**	**58.91**	**50.84**	**35.01**	**50.40**	58.75	45.03	**59.72**

Strategy	Setting	Qwen-3-14B			Ministral-3-14B-Instruct-2512			DeepSeek-V3.2 <deepseek-chat>		
		Micro-F1(%)	Macro-F1(%)	Example-F1(%)	Micro-F1(%)	Macro-F1(%)	Example-F1(%)	Micro-F1(%)	Macro-F1(%)	Example-F1(%)
Label-wise		48.25	**44.33**	48.57	43.31	33.84	43.58	55.26	**49.75**	55.13
	+ Context	48.60 ↑	42.09 ↓	48.70 ↑	43.96 ↑	34.52 ↑	43.85 ↑	56.33 ↑	47.49 ↓	56.21 ↑
List-wise		47.94	43.73	48.65	41.63	32.01	41.63	49.85	41.57	49.97
	+ Context	49.82 ↑	43.10 ↓	50.92 ↑	38.92 ↓	30.48 ↓	38.75 ↓	51.34 ↑	43.82 ↑	50.82 ↑
Path-wise		45.40	39.66	44.92	44.82	34.35	44.65	49.55	43.38	48.57
	+ Context	46.88 ↑	40.26 ↑	46.73 ↑	46.71 ↑	**36.76** ↑	46.47 ↑	55.69 ↑	48.54 ↑	55.72 ↑
OPINE		**54.19**	43.80	**54.43**	**48.43**	34.21	**48.89**	**58.49**	46.29	**58.13**

Table 4: Comparison of scientific opinion classification performance across multiple LLMs under different scoring strategies and the OPINE framework. The best result is **bold** and the second best is underlined. ↑means the improvement of over the baseline configuration that uses only sentence content information; while ↓means the decrease. OPINE is built upon the context-enhanced settings .

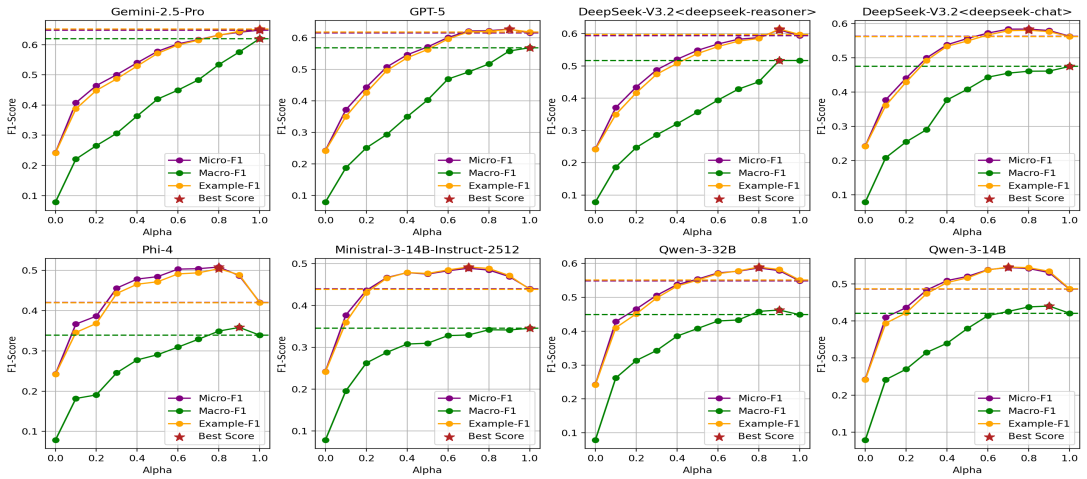

Figure 4: Sensitivity analysis of α under the context-enhanced setting. $\alpha = 0$ corresponds to the AIMRaD prior (selecting the label with the highest probability per section), while $\alpha = 1$ uses affinity scores only (dashed lines). Red stars mark peak performance.

32B (Table 3), we observe that OPINE consistently provides substantial gains. Notably, the performance boost is particularly pronounced on the smaller architecture: Qwen-3-14B achieves a 5.59-point Micro-F1 jump (↑5.59%, 48.60% → 54.19%) under our framework that significantly exceeds the baseline results of its larger 32B counterpart (↑3.82%, 54.85% → 58.67%). This suggests that while larger models have stronger inherent reasoning, structured priors act as a "logical scaffold" that more significantly compensates for the scale-related limitations of smaller architectures. Consequently, OPINE enables mid-sized models to bridge the performance gap, achieving high-level results with far greater parameter efficiency.

5.3 Further Analysis

Superiority of Label-wise Scoring. To determine the optimal backbone for our framework, we compare three scoring strategies: *Label-wise*, *List-wise*, and *Path-wise*. As illustrated in Table 4, the *Label-wise* strategy generally yields the most stable and superior performance across various LLMs. Taking Phi-4 as an example, the *Label-wise* strategy achieves a Micro-F1 of 39.09%, notably exceeding *List-wise* (36.89%) and *Path-wise* (33.89%) approaches. This superiority likely stems from reducing the model's cognitive load by evaluating labels independently, thereby avoiding cumulative errors from complex paths. Consequently, we adopt *Label-wise* scoring as the backbone of OPINE.

Model	Method	Micro			Macro			Example		
		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Qwen-3-14B	**OPINE**	**51.30**	**57.43**	**54.19**	**53.36**	45.96	43.80	**53.38**	**57.96**	**54.43**
	w/o Prior	43.95	55.33	48.99	48.21	**49.37**	43.37	47.46	55.56	49.73
	w/o Refine	51.26	**57.43**	54.17	53.27	45.93	**43.81**	53.34	**57.96**	54.40
	w/o All	46.01	51.50	48.60	46.91	46.09	42.09	47.74	51.95	48.70
Phi-4	**OPINE**	**48.13**	**53.87**	**50.84**	**41.04**	37.50	**35.01**	**48.81**	**54.66**	**50.40**
	w/o Prior	38.53	47.31	42.47	39.61	**37.51**	34.18	38.70	47.60	41.37
	w/o Refine	48.09	**53.87**	50.82	40.97	37.41	34.91	**48.81**	**54.66**	**50.40**
	w/o All	39.79	44.58	42.05	38.37	36.81	33.89	41.09	44.98	41.94
DeepSeek-V3.2 <deepseek-chat>	**OPINE**	**55.37**	**61.99**	**58.49**	**52.97**	48.91	46.29	**56.43**	**62.49**	**58.13**
	w/o Prior	52.73	59.89	56.08	52.13	**50.53**	47.00	54.53	60.33	56.02
	w/o Refine	55.29	61.90	58.41	52.78	48.80	46.09	56.39	62.44	58.07
	w/o All	53.25	59.80	56.33	52.40	**50.53**	**47.49**	54.86	60.28	56.21

Table 5: Ablation study of **Prior Fusion** and **Pairwise Refinement** in OPINE. "Prior" and "Refine" denote the AIMRaD-aware prior fusion and the refinement for confusable labels, respectively.

Sensitivity Analysis of α. We evaluate the fusion weight α by varying it from 0.1 to 1.0 with a step size of 0.1 across *standalone* (Fig. 5) and *context-enhanced* (Fig. 4) settings. Performance curves consistently exhibit an "inverted U-shape," with peak F1-scores concentrated in the $\alpha \in [0.6, 0.9]$ interval. This confirms that neither the *AIMRaD prior* ($\alpha \to 0$) nor *affinity scores* ($\alpha = 1$) alone suffice. Notably, while context enhancement raises the overall performance floor, the structural prior α remains essential for error correction, especially when LLMs generate "hallucinated" labels that deviate from the paper's logical flow. The robustness of this optimal range across diverse architectures demonstrates OPINE's effectiveness in balancing local semantics with global structural constraints.

5.4 Ablation Studies

We conduct ablation experiments to verify the contributions of OPINE components (Table 5). Specifically, we compare the full model with three variants: (1) *w/o Prior* (removes prior fusion), (2) *w/o Refine* (removes pairwise refinement), and (3) *w/o All* (base model). Results show that both modules are essential. "Prior" provides the largest boost (e.g., 5.20% Micro-F1 for Qwen-3-14B) by anchoring predictions within the logical flow. Notably, "Refine" yields a robust independent gain: even without structural priors, the *w/o Prior* variant outperforms the base model (*w/o All*) by 0.39%, confirming its ability to resolve semantic ambiguities. The synergy between global anchoring and pairwise discrimination yields optimal performance.

6 Conclusion

In this paper, we focus on scientific opinion classification based on discourse functions, a task essential for structural reasoning in scientific discourse. To address the inherent multi-label nature of these opinions and the constraints of the AIMRaD structure, we propose OPINE, a multi-stage framework that shifts the paradigm from direct label generation to a controllable *scoring-calibration-refinement* pipeline. By decoupling textual evidence from decision logic, OPINE produces independent affinity scores that are calibrated with explicit discourse priors. Our introduction of a quantile-based decoding rule effectively enables the model to capture co-existing discourse functions, while the pairwise refinement mechanism successfully mitigates confusion between functionally similar categories. Experimental results on our new benchmark of 18 roles demonstrate that OPINE significantly outperforms strong baselines. Our findings reveal that integrating discourse structures as explicit priors is far more effective than conventional passive context integration, offering a more robust approach for nuanced scientific text analysis.

Limitations

Despite the effectiveness of OPINE, our work has several limitations. First, while the quantile-based decoding rule successfully handles multi-label assignment by capturing co-existing roles, the optimal quantile threshold is currently determined through a data-driven heuristic on the validation set. Future work could explore more dynamic and instance-level thresholding mechanisms to better adapt to varying degrees of opinion complexity. Second, although integrating AIMRaD structural priors significantly improves performance, our framework currently assumes a standard discourse structure. Its generalizability to non-standard scientific formats, such as short workshop papers or

interdisciplinary reports with non-linear sections, remains to be further validated in future studies.

Acknowledgments

The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

Waleed Abdeen, Michael Unterkalmsteiner, Krzysztof Wnuk, Alessio Ferrari, and Panagiota Chatzipetrou. 2025. Language models to support multi-label classification of industrial data. In *2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 45–55.

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir R. Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 596–606.

Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G. Iyer, Mugdha S. Polimera, Michael J. Smith, Tirthankar Ghosal, Marc Huertas-Company, Sandor Kruk, Kevin Schawinski, and Ioana Ciucă. 2025. A survey on hypothesis generation for scientific discovery in the era of large language models. *arXiv preprint arXiv:2504.05496*.

Margaret Cargill and Patrick D. T. O'Connor. 2013. *Writing Scientific Research Articles: Strategy and Steps*. John Wiley & Sons.

Yi-Pei Chen, KuanChao Chu, and Hideki Nakayama. 2024. Llm as a scorer: The impact of output order on dialogue evaluation. *arXiv preprint arXiv:2406.02863*.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 2489–2513.

Andrew Curtis. 2012. The science of subjectivity. *Geology*, 40(1):95–96.

Wenlong Hu, Qiang Fan, Hao Yan, Xinyao Xu, Shan Huang, and Ke Zhang. 2025. A survey of multi-label text classification under few-shot scenarios. *Applied Sciences*, 15(16):8872.

Edward J. Huth. 1987. Structured abstracts for papers reporting clinical trials. *Annals of Internal Medicine*, 106(4):626–627.

Kripa Jagannathan, Geniffer Emmanuel, James Arnott, Katharine J. Mach, Aparna Bamzai-Dodson, Kristen Goodrich, Ryan Meyer, Mark Neff, K. Dana Sjostrom, Kristin M. F. Timm, Esther Turnhout, Gabrielle Wong-Parodi, Angela T. Bednarek, Alison Meadow, Art Dewulf, Christine J. Kirchhoff, Richard H. Moss, Leah Nichols, Eliza Oldach, and 2 others. 2023. A research agenda for the science of actionable knowledge: Drawing from a review of the most misguided to the most enlightened claims in the science–policy interface literature. *Environmental Science & Policy*, 144:174–186.

Noriko Kando. 1997. Text-level structure of research papers: Implications for text-based information processing systems. In *Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research*. BCS Learning & Development.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Jieran Li, Xiuyuan Hu, Yang Zhao, Shengyao Zhuang, and Hao Zhang. 2025. Leveraging reference documents for zero-shot ranking via large language models. *arXiv preprint arXiv:2506.11452*.

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. 2025. Large language models do multi-label classification differently. *arXiv preprint arXiv:2505.17510*.

Michael J. Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science*, 5(1):86–92.

Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leipold. 2024. Afacta: Assisting the annotation of factual claim detection with reliable LLM annotators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1890–1912.

Mohammad Rabiei, Seyyed-Mahdi Hosseini-Motlagh, and Abdorrahman Haeri. 2017. Using text mining techniques for identifying research gaps and priorities: a case study of the environmental science in iran. *Scientometrics*, 110(2):815–842.

Wanda Safira, Benedictus Prabaswara, Andrea Stevens Karnyoto, and Bens Pardamean. 2025. Leveraging ALBERT for sentiment classification of long-form ChatGPT reviews on twitter. *International Journal of Computing and Digital Systems*, 17(1):1–12.

Hajar Sakai and Sarah S Lam. 2025. Quad-llm-mltc: Large language models ensemble learning for healthcare text multi-label classification. *arXiv preprint arXiv:2502.14189*.

Ingeborg Spiegel-Rüsing. 1977. Science studies: Bibliometric and content analysis. *Social Studies of Science*, 7(1):97–113.

Shuoqi Sun, Shengyao Zhuang, Shuai Wang, and Guido Zuccon. 2025. An investigation of prompt variations for zero-shot llm-based rankers. In *European Conference on Information Retrieval*, pages 185–201. Springer.

John Swales. 2004. *Research Genres: Explorations and Applications*. Cambridge University Press.

John Swales. 2014. Create a research space (cars) model of research introductions. In *Writing About Writing: A College Reader*, pages 12–15.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Phd thesis, University of Edinburgh, Edinburgh, UK.

Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. Center for the Study of Language and Information.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Mingxuan Xia, Zhijie Jiang, Haobo Wang, Junbo Zhao, Tianlei Hu, and Gang Chen. 2025. Ensembling prompting strategies for zero-shot hierarchical text classification with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.

Ruichao Yang, Jing Ma, Wei Gao, and Hongzhan Lin. 2025. LLM-enhanced multiple instance learning for joint rumor and stance detection with social context information. *ACM Transactions on Intelligent Systems and Technology*, 16(3):1–27.

Kosuke Yoshimura and Hisashi Kashima. 2025. Hierarchical text classification using black box large language models. *arXiv preprint arXiv:2508.04219*.

Jinze Yu and Guanghui Wang. 2025. Reasoning-enhanced prompt strategies for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6065–6070.

Mengting Zhang, Yajiao Wang, Yufei Wang, and Zhixiong Zhang. 2025a. Sentiment-enhanced opinion sentence recognition using AgriBERT-SentiDPCNN and multi-opinion summarization via LLMs for agricultural scientific literature. In *Proceedings of the 10th International Conference on Big Data Analytics*, pages 300–308. IEEE.

Mengting Zhang, Zhixiong Zhang, Yajiao Wang, Yang Li, Xin Lin, and Meng Wang. 2025b. Explain before classify: Contrastive rationale distillation for academic opinion recognition. In *Proceedings of the International Conference on Advanced Data Mining and Applications*, pages 311–325, Singapore. Springer Nature Singapore.

Qian Zhang, Qinliang Su, Wei Zhu, and Yachun Pang. 2025c. Hierprompt: Zero-shot hierarchical text classification with llm-enhanced prototypes. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3846–3859.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2025d. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. In *Proceedings of the ACM on Web Conference 2025 (WWW '25)*, pages 2032–2042.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yaxin Zhu and Hamed Zamani. 2024. Icxml: An in-context learning framework for zero-shot extreme multi-label classification. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2086–2098.

Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. Can large language models understand context? *arXiv preprint arXiv:2402.00858*.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2023. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. *arXiv preprint arXiv:2310.14122*.

A More Related Works

Prior work on opinion classification has primarily focused on user-generated content such as social media and product reviews (Wankhade et al., 2022), constructing taxonomies around sentiment polarity (Safira et al., 2025) or stance (Yang et al., 2025) that are ill-suited to the evidence-based, reasoning-driven judgments prevalent in scientific writing. To analyze scientific texts, Teufel(Teufel, 1999) proposed *Argumentative Zoning* to assign functional roles to sentences based on their rhetorical purpose, leveraging the structured organization of scientific articles such as IMRaD (Huth, 1987). Swales(Swales, 2004) further modeled scientific discourse through *rhetorical moves* and introduced the CARS framework(Swales, 2014) to characterize the argumentative structure of research articles. In parallel, citation analysis studies categorize citation sentences according to their functional roles in scientific discourse, including purpose, polarity, and usage patterns(Moravcsik and Murugesan, 1975; Spiegel-Rüsing, 1977; Teufel et al., 2006; Abu-Jbara et al., 2013). These works analyze sentences in terms of their discourse functions within scientific argumentation, which is closely related to our goal of modeling the functional types of opinion-bearing sentences in scientific texts.

B Additional Experimental Results and Analysis

B.1 Additional Experimental Results

Tables 6, 7, and 8 list the detailed Micro, Macro, and Example-based metrics for Qwen-3-14B, Phi-4, and Ministral-3-14B-Instruct-2512, respectively. Complementing the sensitivity analysis in Fig. 4, we provide Fig. 5 to illustrate the behavior of the fusion weight α in a standalone setting (where surrounding text context is removed).

B.2 Theoretical Rationale: Narrative Mandates and Opinion Distribution

The distribution of scientific opinions is intrinsically shaped by the functional requirements of research discourse. Each section serves as a specific epistemic stage in the overall scientific argument:

- **Introduction**: Background Contextualization and Niche Identification. The *Introduction* follows a progressive structure, moving from broad field-level overviews to specific research focuses. By reviewing existing literature, authors define the "knowledge frontier" and identify a research niche—an unresolved gap that justifies the current study. Consequently, opinions in this section are primarily evaluative, centering on the ***significance of the domain***, ***limitations of prior works***, and the ***necessity of the proposed research***.

- **Materials and methods**: Establishing Methodological Trustworthiness. The core role of the *Materials and methods* section is to establish the credibility of the findings by detailing how the research was conducted. Beyond simple reproducibility, its fundamental goal is to allow readers and reviewers to assess whether the research design and analytical procedures adhere to rigorous scientific norms. Authorial opinions here are thus focused on ***justifying methodological choices***, the ***rationality of experimental steps***, and the ***standardization of data analysis***.

- **Results**: Empirical Interpretation and Evidence. Positioned at the heart of the research narrative, the *Results* section presents the empirical evidence used to support or refute hypotheses. Since this section is strictly anchored in empirical materials, opinions expressed here typically derive from the data itself, involving the ***interpretation of observations***, ***evaluation of findings' significance***, and ***comparative analysis of differences across experimental conditions***.

- **Discussion**: Knowledge Integration and Higher-Level Inference. The *Discussion* raises the focus from specific evidence back to the macro-perspective of the field, showing how the findings integrate into a broader knowledge framework. It summarizes key findings in relation to the original hypotheses and evaluates their alignment with existing literature. As the most opinion-dense portion of the paper, it does not repeat raw data but provides high-level inferences, ***acknowledges study limitations***, and ***extrapolates theoretical or practical implications***.

B.3 Taxonomy and Formal Definitions

Following the narrative tasks identified in Appendix B.2, we define 18 fine-grained scientific opinion types (see Table 1). These categories ensure that each authorial judgment is mapped to a

Paradigm	Strategy	Setting	Micro			Macro			Example		
			P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
LLMs as Label Generators	Flattened		13.40	77.39	22.85	12.39	77.94	20.20	17.40	78.07	27.04
		+ Context	12.72 ↓	77.21 ↓	21.84 ↓	12.02 ↓	76.45 ↓	19.46 ↓	16.20 ↓	77.60 ↓	25.62 ↓
		+ AIMRaD	13.71 ↑	**81.59** ↑	23.47 ↑	13.00 ↑	**81.75** ↑	21.08 ↑	17.41 ↑	**82.11** ↑	27.47 ↑
		+ All	13.55 ↑	79.03 ↑	23.13 ↑	12.89 ↑	79.50 ↑	20.79 ↑	16.89 ↓	79.35 ↑	26.69 ↓
	Per-parent		40.48	32.54	36.08	48.64	31.15	32.28	31.24	32.71	31.51
		+ Context	44.82 ↑	33.91 ↑	38.61 ↑	44.97 ↓	29.03 ↓	29.48 ↓	32.62 ↑	34.02 ↑	32.78 ↑
		+ AIMRaD	47.34 ↑	36.46 ↑	41.19 ↑	41.66 ↓	31.83 ↑	30.94 ↓	34.67 ↑	36.89 ↑	35.18 ↑
		+ All	**52.36** ↑	36.46 ↑	42.99 ↑	48.87 ↑	31.96 ↑	33.59 ↑	35.38 ↑	36.73 ↑	35.51 ↑
	Per-path		21.06	73.56	32.75	23.83	68.52	31.76	26.22	73.71	36.93
		+ Context	20.85 ↓	74.20 ↑	32.55 ↓	22.94 ↓	68.55 ↑	30.59 ↓	25.63 ↓	74.42 ↑	36.27 ↓
		+ AIMRaD	20.54 ↓	76.39 ↑	32.38 ↓	21.44 ↓	73.00 ↑	30.05 ↓	25.57 ↓	76.45 ↑	36.56 ↓
		+ All	22.60 ↑	76.48 ↑	34.89 ↑	24.45 ↑	70.11 ↑	32.23 ↑	27.76 ↑	77.01 ↑	38.91 ↑
LLMs as Label Scorers	Label-wise		45.38	51.50	48.25	49.39	47.98	**44.33**	47.47	52.03	48.57
		+ Context	46.01 ↑	51.50	48.60 ↑	46.91 ↑	46.09 ↓	42.09 ↓	47.74 ↑	51.95 ↓	48.70 ↑
	List-wise		43.31	53.69	47.94	51.71	49.62	43.73	46.43	53.88	48.63
		+ Context	44.37 ↑	56.79 ↑	49.82 ↑	45.63 ↓	49.69 ↑	43.10 ↓	48.24 ↑	57.40 ↑	50.92 ↑
	Path-wise		42.96	48.13	45.40	49.44	42.41	38.66	43.61	48.60	44.92
		+ Context	44.38 ↑	49.68 ↑	46.88 ↑	47.00 ↓	42.90 ↑	40.26 ↑	45.57 ↑	50.37 ↑	46.73 ↑
	OPINE		51.30	57.43	**54.19**	**53.36**	45.96	43.80	**53.38**	57.96	**54.43**

Table 6: Main results of different LLM-based paradigms under various configurations (Qwen-3-14B). The best result is **bold** and the second best is underlined. ↑means the improvement of over the baseline configuration that uses only sentence content information; while ↓means the decrease. OPINE is built upon the context-enhanced settings .

specific discourse function, facilitating consistent analysis across diverse scientific corpora.

B.3.1 A. Scientific Community

A.1 Research Actors

- **A.1-1 Opinions on the contribution and influence of research actors**
  Definition: Evaluations regarding the pioneering nature, innovation, or academic impact of specific researchers or research teams.
  Example: "Li et al. **pioneered the use** of the MRC framework to handle both flat and nested entities."

A.2 Research Outputs

- **A.2-1 Opinions on the value and impact of research outputs**
  Definition: Assessment of the effectiveness, utility, or performance of specific products like models, datasets, or software.
  Example: "BERT is a classic pre-training model that has shown **great effectiveness** in various tasks."

A.3 Research Objects

- **A.3-1 Overall status and development stage**
  Definition: Judgments on the macro-level trends or the current evolutionary phase (e.g., emerging, mature) of a field.
  Example: "Current research on multi-modal metaphor detection is still **in its early stages**."

- **A.3-2 Current key research priorities and hotspots**
  Definition: Identification of topics that currently dominate the field's attention due to their importance or urgency.
  Example: "Understanding semantics in multimodal utterances **has attracted much attention** with the boom..."

- **A.3-3 Mainstream knowledge and consensus**
  Definition: Summarizing beliefs or findings that have been widely validated and accepted by the scientific community.
  Example: "**It is known that** the effective design of task-specific prompts is critical for LLMs."

- **A.3-4 Mainstream technical routes**
  Definition: Identifying the standard methodologies or dominant strategies currently used to solve specific problems.
  Example: "Weight quantization has **emerged**

Paradigm	Strategy	Setting	Micro			Macro			Example		
			P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
LLMs as Label Generators	Flattened		30.83	63.17	41.43	30.50	58.75	**36.44**	32.72	63.27	42.35
		+ Context	28.08 ↓	63.35 ↑	38.91 ↓	27.25 ↓	55.76 ↓	32.97 ↓	29.71 ↓	63.89 ↑	39.92 ↓
		+ AIMRaD	31.15 ↑	**63.45** ↑	41.79 ↑	29.14 ↓	**59.38** ↑	35.98 ↓	33.56 ↑	**63.95** ↑	43.15 ↑
		+ All	28.47 ↓	63.08 ↓	39.23 ↓	28.58 ↓	55.26 ↓	33.34 ↓	30.17 ↓	63.74 ↑	40.31 ↓
	Per-parent		27.96	24.34	26.02	33.94	23.73	18.74	18.80	24.44	20.48
		+ Context	25.32 ↓	21.97 ↓	23.52 ↓	28.94 ↓	20.55 ↓	16.35 ↓	15.82 ↓	22.21 ↓	17.68 ↓
		+ AIMRaD	32.93 ↑	27.53 ↑	29.99 ↑	31.87 ↓	26.72 ↑	21.63 ↑	21.31 ↑	27.77 ↑	23.20 ↑
		+ All	26.98 ↓	26.07 ↑	26.52 ↑	30.78 ↓	23.12 ↓	17.25 ↓	18.74 ↓	26.44 ↑	21.03 ↑
	Per-path		28.66	58.16	38.40	30.47	51.25	31.09	33.34	58.15	40.85
		+ Context	28.32 ↓	61.44 ↑	38.77 ↑	37.77 ↑	53.57 ↑	32.65 ↑	30.95 ↓	61.58 ↑	40.22 ↓
		+ AIMRaD	31.21 ↑	61.71 ↑	41.46 ↑	35.28 ↑	52.80 ↑	33.22 ↑	36.45 ↑	61.96 ↑	44.30 ↑
		+ All	27.57 ↓	59.80 ↑	37.74 ↓	31.86 ↑	51.99 ↑	31.63 ↑	30.61 ↓	59.96 ↑	39.49 ↓
LLMs as Label Scorers	Label-wise		36.83	41.66	39.09	**47.49**	35.73	33.91	36.27	41.79	37.72
		+ Context	39.79 ↑	44.58 ↑	42.05 ↑	38.37 ↓	36.81 ↑	33.89 ↓	41.09 ↑	44.98 ↑	41.94 ↑
	List-wise		32.88	42.02	36.89	36.22	35.99	30.84	32.94	42.17	35.75
		+ Context	33.18 ↑	40.20 ↓	36.36 ↓	31.44 ↓	34.35 ↓	28.81 ↓	32.84 ↓	40.23 ↓	35.07 ↓
	Path-wise		32.08	35.92	33.89	34.15	26.96	24.62	30.55	36.30	32.24
		+ Context	31.81 ↓	35.64 ↓	33.62 ↓	35.34 ↑	28.42 ↑	25.90 ↑	31.08 ↑	35.73 ↓	32.35 ↑
	OPINE		**48.13**	53.87	**50.84**	41.04	37.50	35.01	**48.81**	54.66	**50.40**

Table 7: Main results of different LLM-based paradigms under various configurations (Phi-4). The best result is **bold** and the second best is underlined. ↑means the improvement of over the baseline configuration that uses only sentence content information; while ↓means the decrease. OPINE is built upon the context-enhanced settings .

as a popular strategy to enhance the efficiency of LLMs."

- **A.3-5 Current research gaps and insufficiencies**
  Definition: Highlighting what remains unexplored or where existing studies fail to provide sufficient solutions.
  Example: "**Unfortunately**, fair evaluations of PPLMs are **still unexplored**."

- **A.3-6 Current research difficulties and challenges**
  Definition: Pointing out inherent technical bottlenecks or task-specific obstacles that are difficult to overcome.
  Example: "Even with the side network, the inherent model size of the LLM **remains a challenge**."

- **A.3-7 Selection and formulation of research problems**
  Definition: The precise definition of the specific technical problem the author intends to address in the study.
  Example: "Quantization can **lead to accuracy degradation**, **attributable to** the inherent information loss."

- **A.3-8 Value of a research field, topic, or specific problem**
  Definition: Arguing for the necessity and significance of the study by highlighting its theoretical or practical impact.
  Example: "NER **plays a crucial role** in facilitating various downstream tasks such as relation extraction."

B.3.2 B. Authors' Own Research

B.1 Research Hypotheses

- **B.1-1 Opinions concerning research hypotheses**
  Definition: Speculative claims or predictions about mechanisms or relationships that the study aims to verify.
  Example: "We **hypothesize** that performance is correlated with the size of the dataset."

B.2 Research Methods

- **B.2-1 Soundness and rationale of methodological design**
  Definition: The subjective justification explaining why a particular method was designed or chosen.
  Example: "**Considering that** LLMs can use their own knowledge, we introduce a **novel perspective**..."

Paradigm	Strategy	Setting	Micro			Macro			Example		
			P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
LLMs as Label Generators	Flattened		36.67	60.44	45.65	38.36	56.62	**41.99**	43.06	60.53	48.37
		+ Context	31.87 ↓	63.99 ↑	42.55 ↓	33.29 ↓	57.40 ↑	36.95 ↓	38.37 ↓	64.16 ↑	45.86 ↓
		+ AIMRaD	37.02 ↑	61.35 ↑	46.17 ↑	39.22 ↑	56.20 ↓	41.47 ↓	43.20 ↑	61.53 ↑	48.72 ↑
		+ All	32.50 ↓	64.90 ↑	43.31 ↓	32.39 ↓	58.89 ↑	37.48 ↓	39.78 ↓	65.16 ↑	47.06 ↓
	Per-parent		63.68	12.31	20.63	28.91	6.51	9.57	11.58	12.41	11.77
		+ Context	59.09 ↓	4.74 ↓	8.78 ↓	26.86 ↓	2.55 ↓	4.46 ↓	4.24 ↓	4.77 ↓	4.39 ↓
		+ AIMRaD	**69.06** ↑	11.39 ↓	19.56 ↓	31.68 ↑	5.87 ↓	8.90 ↓	11.04 ↓	11.53 ↓	11.09 ↓
		+ All	59.78 ↓	5.01 ↓	9.25 ↓	26.05 ↓	2.43 ↓	4.22 ↓	4.39 ↓	4.97 ↓	4.53 ↓
	Per-path		28.04	53.60	36.82	35.23	53.61	35.39	28.13	53.64	35.71
		+ Context	26.27 ↓	62.44 ↑	36.98 ↑	29.05 ↓	61.59 ↑	35.40 ↑	29.28 ↑	62.74 ↑	38.65 ↑
		+ AIMRaD	30.92 ↑	60.35 ↑	40.89 ↑	36.06 ↑	59.85 ↑	39.38 ↑	33.17 ↑	60.17 ↑	41.17 ↑
		+ All	28.28 ↑	**65.27** ↑	39.46 ↑	30.57 ↓	**64.01** ↑	37.17 ↑	32.00 ↑	**65.48** ↑	41.51 ↑
LLMs as Label Scorers	label-wise		40.89	46.03	43.31	37.97	36.99	33.84	42.85	46.46	43.58
		+ Context	41.61 ↑	46.58 ↑	43.96 ↑	39.02 ↑	37.86 ↑	34.52 ↑	42.95 ↑	46.98 ↑	43.85 ↑
	List-wise		37.34	47.04	41.63	31.68	38.33	32.01	39.42	47.36	41.63
		+ Context	34.98 ↓	43.85 ↓	38.92 ↓	33.49 ↑	35.82 ↓	30.48 ↓	36.54 ↓	44.06 ↓	38.75 ↓
	Path-wise		42.43	47.49	44.82	38.95	38.69	34.35	43.71	47.98	44.65
		+ Context	44.22 ↑	49.50 ↑	46.71↑	**44.59** ↑	40.46 ↑	36.76 ↑	45.42 ↑	49.84 ↑	46.47 ↑
	OPINE		45.85	51.32	**48.43**	38.84	36.95	34.21	**48.05**	51.89	**48.89**

Table 8: Main results of different LLM-based paradigms under various configurations (Ministral-3-14B-Instruct-2512). The best result is **bold** and the second best is underlined. ↑means the improvement of over the baseline configuration that uses only sentence content information; while ↓means the decrease. OPINE is built upon the context-enhanced settings .

B.3 Experimental Design

- **B.3-1 Soundness and rationale of experimental design**
  Definition: Explanation of the logic behind the experimental setup, data selection, or baseline comparisons.
  Example: "The **rationale for using** these datasets is that they exhibit a variety of distinct semantics."

B.4 Research Results

- **B.4-1 Interpret research results or provide causal explanations**
  Definition: Inferring the meaning of experimental outcomes or attributing results to specific causes.
  Example: "The severe performance drop of DetectGPT is **attributed to** its reliance on accessibility..."

- **B.4-2 Compare research results and judge superiority or effectiveness**
  Definition: Asserting the comparative advantage or SOTA performance of the proposed method based on data.
  Example: "AoE achieved **SOTA performance** in BERT-large scale models, with an average score of 64.24."

B.5 Research Conclusions

- **B.5-1 Contributions and value of the authors' own research**
  Definition: Defining the novel increment and theoretical/practical value the study adds to the field.
  Example: "In this work, we make a **pioneering contribution** by formulating the discovery task."

- **B.5-2 Limitations and insufficiencies of the authors' own research**
  Definition: Self-reflective assessment of the boundaries, weaknesses, or constraints of the current work.
  Example: "The **main limitation** is that due to the **lack of computing resources**, we only used 7B models."

- **B.5-3 Future improvements and development directions**
  Definition: Projecting future research paths or specific optimization strategies for the current study.
  Example: "Our **future work** will **focus on**

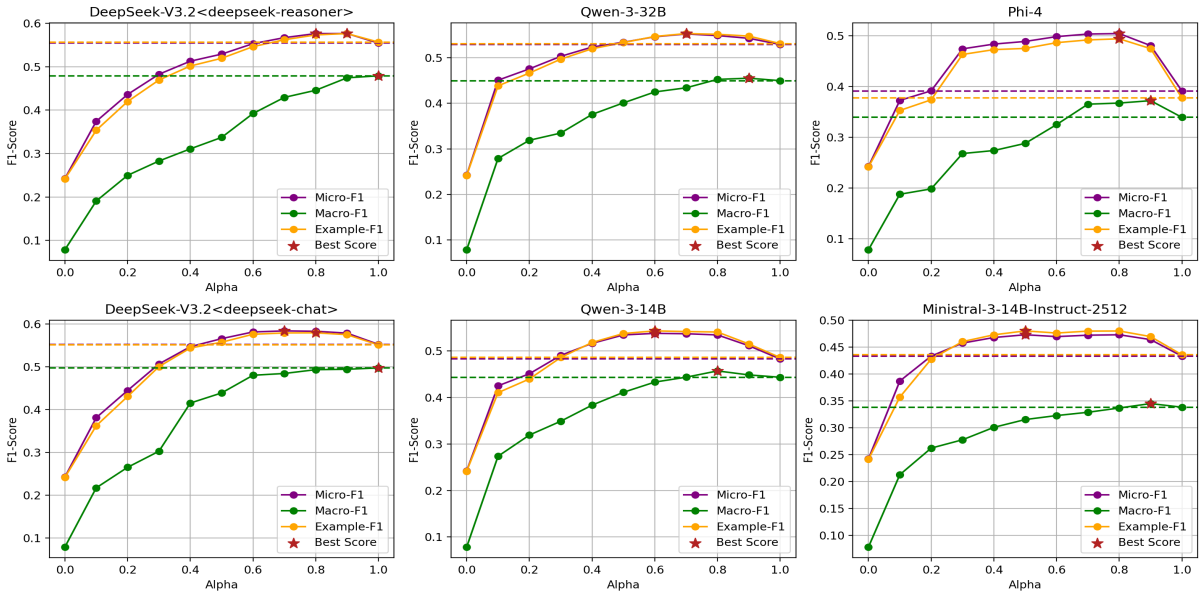

Figure 5: Sensitivity analysis of α in the standalone setting (excluding context). $\alpha = 0$ corresponds to the AIMRaD prior (selecting the label with the highest probability per section), while $\alpha = 1$ uses affinity scores only (dashed lines). Red stars mark peak performance.

how to better engage LLMs into the translation part."

C Derivation of the AIMRaD-Aware Prior Fusion

We aim to estimate the posterior probability $P(y \mid x, c, s)$, where $y \in \mathcal{Y}$ is the opinion label, x is the target sentence, c is the local context, and s is the AIMRaD section. Following Bayes' Theorem:

$$P(y \mid x, c, s) = \frac{P(x, c, s \mid y)\, P(y)}{P(x, c, s)} \propto P(x, c, s \mid y)\, P(y) \tag{8}$$

To make the computation tractable, we introduce a conditional independence assumption: given the specific functional label y, the linguistic realization (textual content x, c) and the structural location (section s) are independent. Intuitively, this assumption implies that once the functional intent of a sentence (e.g., "stating a hypothesis") is determined, the specific linguistic phrasing used to express that intent is independent of which section the sentence resides in. While the prior distribution of labels y varies significantly across sections s, the generative process of the text (x, c) is driven primarily by the label's semantic requirements rather than the section's structural index. Formally:

$$P(x, c, s \mid y) \approx P(x, c \mid y) P(s \mid y) \tag{9}$$

Substituting this into the Bayesian formulation:

$$P(y \mid x, c, s) \propto P(x, c \mid y) P(s \mid y) P(y) \tag{10}$$

By applying Bayes' rule again to the individual terms:

- $P(x, c \mid y) \propto \frac{P(y|x,c)}{P(y)}$ (Text-based likelihood)
- $P(s \mid y) \propto \frac{P(y|s)}{P(y)}$ (Section-based likelihood)

The posterior can then be rewritten as:

$$P(y \mid x, c, s) \propto \frac{P(y \mid x, c) P(y \mid s)}{P(y)} \tag{11}$$

In the theoretically grounded formulation (Eq. 11), $P(y)$ acts as a marginal debiasing factor. Its role is to penalize labels with high global frequency, ensuring that the final prediction is not disproportionately biased toward majority classes that appear frequently across all sections.

However, we transition from this generative derivation to a discriminative log-linear fusion for two practical considerations: **First**, the textual predictor $P_{\text{text}}(y \mid x, c)$, typically parameterized by a Large Language Model (LLM), is not a naive likelihood but an informed posterior estimate. Since the LLM already implicitly internalizes the marginal label distribution $P(y)$ through its pre-training and in-context instructions, explicitly dividing by an empirical $P(y)$ may introduce

redundant noise. **Second**, to adaptively control the influence of discourse-level constraints across diverse scientific domains, we generalize the product into a weighted geometric mean. In this parameterized setup, the effect of the marginal prior $P(y)$ is effectively absorbed into the tunable balancing parameter α and the final normalization constant $Z(x, c, s)$. This yields the final posterior objective utilized in our methodology:

$$P^*(y \mid x, c, s) = \frac{P_{\text{text}}(y \mid x, c)^{\alpha} P_{\text{prior}}(y \mid s)^{1-\alpha}}{\sum_{y' \in \mathcal{Y}} P_{\text{text}}(y' \mid x, c)^{\alpha} P_{\text{prior}}(y' \mid s)^{1-\alpha}} \quad (12)$$

where $\alpha \in (0, 1]$ serves as a hyperparameter to balance textual evidence and structural priors. This formulation maintains the theoretical essence of Bayesian calibration while providing the empirical robustness necessary for fine-grained scientific discourse analysis.

D More Details about the Dataset

D.1 Annotation Guidelines and Procedures

To ensure high-quality and consistent labels for the 18 discourse functions, we developed a comprehensive annotation protocol. This section details the instructions provided to annotators, the category definitions, and the quality control mechanisms employed.

Annotators were provided with a web-based interface and a annotation manual. The core task was to identify the functional role(s) of each sentence within its scientific context. Key instructions included:

- **Utilization of AIMRaD Priors:** Annotators were encouraged to use the sentence's position within the AIMRaD structure as a structural prior. For example, a "Hypothesis" is more likely to appear in the Introduction or Methods than in the Results section.
- **Multi-label Assignment:** Given the complexity of scientific discourse, sentences may serve multiple purposes. Annotators were permitted to assign multiple labels if a sentence explicitly performed several functions (e.g., both a *Result* and a *Comparison*).
- **Contextual Reading:** To capture long-range dependencies, annotators were required to read at least two sentences of the preceding and following context before assigning a label.

All annotators were briefed on the nature of the research, the specific tasks involved, and the intended use of the resulting dataset for academic publication.

D.2 Section Mapping to AIMRaD

To facilitate a standardized analysis across diverse paper formats, we mapped the original section titles from the dataset into the **AIMRaD** structural framework. This mapping was performed using a priority-based keyword matching heuristic:

- **Abstract**: Identifies sections explicitly labeled as *abstract*.
- **Introduction**: Includes *introduction, background, related work, motivation, preliminaries,* and *problem statement*.
- **Materials and method**: Includes keywords such as *methodology, approach, architecture, framework, algorithm, system,* and *implementation details*.
- **Result**: Includes *experiments, evaluation, results, empirical analysis, ablation study,* and *case studies*.
- **Discussion**: Includes keywords such as *discussion, conclusion, limitations, future work,* and *outlook*.

  Note: For sections that did not match any specific keywords (e.g., specific technical headings or model names), we defaulted the classification to ***Materials and method****, as these segments typically describe the core technical contribution and proposed system of the paper.*

D.3 Data Format and Field Definitions

The processed dataset is organized in a JSONL (JSON Lines) format. Each entry represents a fine-grained opinion unit extracted from a paper, enriched with structural metadata. The fields are defined as follows:

- `paper_id`: The identifier of the source paper (e.g., from ACL Anthology), often including the paper title for reference.
- `label`: The taxonomic category of the scientific opinion sentence (e.g., `A.3-3`).

- `section_raw`: The original section heading as parsed from the source document (e.g., "2 Related Work").
- `section_norm`: The mapped functional section under the **AIMRaD** framework (Abstract, Introduction, Materials and methods, Result, or Discussion), derived from `section_raw` via our heuristic mapping logic.
- `sentence`: The specific sentence identified as containing a scientific opinion.
- `context`: The full paragraph that provide the necessary background for interpreting the `sentence`.

A representative data instance from our dataset is illustrated below:

```
{
  "paper_id": "2024.acl-long.1(Quantized Side
     Tuning: Fast and Memory-Efficient Tuning of
     Quantized Large Language Models)",
  "label": "B.5-1",
  "section_raw": "Abstract",
  "section_norm": "Abstract",
  "sentence": "In this paper, we present
     quantized side tuning (QST), which enables
     memory-efficient and fast finetuning of
     LLMs by operating through a dual-stage
     process.",
  "context": "Finetuning large language models
     (LLMs) has been... [full text]"
}
```

E Prompt Details

E.1 System Prompt

```
You are an expert assistant for **academic
     opinion classification**.

Task:
Given a sentence from a scientific paper (with
     optional local context),
judge how likely the author's MAIN academic
     opinion in that sentence
belongs to each predefined opinion type.

Taxonomy:
Academic opinions are organised into **
     opinion-object groups** and
**fine-grained labels**.

1. Opinion-object groups (coarse level)

{GROUP_DEF_BLOCK}

2. Fine-grained opinion labels (detailed level)

{OPINION_DEF_BLOCK}

3. Group-label structure

{GROUP_LABEL_BLOCK}

General rules:
- Only the predefined fine-grained labels may
     be used; no new labels may be created.
- A sentence may express one or multiple
     opinion types simultaneously.
- Whenever you produce scores, you must provide
     a score for EVERY fine-grained label.
- Follow the numerical constraints and output
     format specified in the user message.
- Do not explain your reasoning unless
     explicitly asked; focus on clean numeric
     output.
```

E.2 Label Schema Definition

```
# -*- coding: utf-8 -*-
# level-1
GROUP_ID2DESC = {
    "A.1": "Opinions about research actors in
    the scientific community (who contributes
    what to the field)",
    "A.2": "Opinions about research outputs
    (the value and impact of concrete
    scientific products such as papers, models,
    datasets)",
    "A.3": "Opinions about research objects and
    problem space (fields, topics, problem
    formulations, gaps, and challenges)",
    "B.1": "Opinions about research hypotheses
    formulated in this study",
    "B.2": "Opinions about research methods and
    methodological design in this study",
    "B.3": "Opinions about experimental schemes
    and experimental design in this study",
    "B.4": "Opinions about research results
    produced by this study (their meaning,
    causes, and comparative performance)",
    "B.5": "Opinions about the overall
    conclusions of this study, including
    contributions, limitations, and future
    research directions"
}

# level-2
OPINION_ID2DESC = {
    "A.1-1": "Opinions on the contribution and
    influence of research actors (e.g.,
    individual researchers, research groups,
    institutions), highlighting who has driven
    the development of a field and in what way",
    "A.2-1": "Opinions on the value and impact
    of specific research outputs (papers,
    models, datasets, tools, systems), focusing
    on their relative advantages, practical
    usefulness, or pioneering role within a
    method family",
    "A.3-1": "Opinions on the overall status
    and development stage of a research field,
    summarising trends, paradigm shifts, or
    whether the field is in an emerging,
    developing, or mature phase",
    "A.3-2": "Opinions on current key research
    priorities and hotspots, identifying core
    topics that are repeatedly discussed or
    highly active in the literature at a given
    time",
```

```
    "A.3-3": "Opinions on mainstream knowledge
        and consensus regarding a research topic,
        stating what has been repeatedly verified
        and is widely accepted in the community
        about certain methods, assumptions, or
        phenomena",
    "A.3-4": "Opinions on mainstream technical
        routes in the field, highlighting which
        solution strategies or methodological
        frameworks have become standard or dominant
        ways of addressing a problem",
    "A.3-5": "Opinions on current research gaps
        and insufficiencies, pointing out
        under-explored areas, missing aspects, or
        limitations of existing work that leave
        important problems unresolved",
    "A.3-6": "Opinions on current research
        difficulties and challenges, revealing
        inherent obstacles or technical bottlenecks
        that make a task particularly hard to
        solve, even when it is already widely
        studied",
    "A.3-7": "Opinions on the selection and
        formulation of research problems,
        explaining how general needs or gaps are
        concretised into specific core questions,
        technical targets, or mechanisms to be
        investigated, and how the problem is
        delimited",
    "A.3-8": "Opinions on the value of a
        research field, topic, or specific problem,
        judging why it is worth studying in terms
        of theoretical importance or practical
        relevance",
    "B.1-1": "Opinions on research hypotheses,
        understood as reasoned but unverified
        expectations or predictions about variable
        relations, mechanisms, or effects,
        typically expressed with verbs such as
        hypothesise, assume, or posit",
    "B.2-1": "Opinions on the rationality of
        research method design, explaining why a
        particular methodological strategy,
        architecture, or component is appropriate
        for the research goals, and how its
        structure matches the task needs",
    "B.3-1": "Opinions on the rationality of
        experimental scheme design, justifying why
        experiments are organised in a specific
        way, including choices of datasets,
        baselines, control settings, and evaluation
        protocols, and how these choices support
        credible comparison and validation",
    "B.4-1": "Opinions that interpret research
        results or provide causal explanations,
        describing what the observed findings mean
        and why they occur, often using cautious
        epistemic verbs such as indicate, suggest,
        or imply, or causal markers such as due to
        and attributable to",
    "B.4-2": "Opinions that compare research
        results and judge superiority or
        effectiveness, emphasising how a method
        performs better, worse, or comparably to
        baselines or prior work, typically using
        comparative expressions like outperform or
        better than",
    "B.5-1": "Opinions on the contribution and
        value of this study as a whole, clarifying
        what knowledge gaps it fills, which
        technical bottlenecks it breaks, or what
        new perspectives it introduces, often
        marked by terms like novel, pioneering, or
        valuable",
    "B.5-2": "Opinions on the limitations and
        insufficiencies of this study, explicitly
        stating constraints related to data, model
        capacity, experimental settings, task
        assumptions, or resources, and thus
        delineating the scope in which the
        conclusions remain valid",
    "B.5-3": "Opinions on future improvements
        and development directions, proposing
        concrete extensions of the current work
        (e.g., larger data, new models, broader
        scenarios) and/or higher-level future
        directions for the field, often introduced
        by expressions such as future work or
        should be further explored",
}

GROUP2LABELS = {
    "A.1": ["A.1-1"],
    "A.2": ["A.2-1"],
    "A.3": ["A.3-1", "A.3-2", "A.3-3", "A.3-4",
        "A.3-5", "A.3-6", "A.3-7", "A.3-8"],
    "B.1": ["B.1-1"],
    "B.2": ["B.2-1"],
    "B.3": ["B.3-1"],
    "B.4": ["B.4-1", "B.4-2"],
    "B.5": ["B.5-1", "B.5-2", "B.5-3"],
}


OPINION_LABELS = sorted(OPINION_ID2DESC.keys())
NUM_LABELS = len(OPINION_LABELS)

LABEL2GROUP = {}
for g, labs in GROUP2LABELS.items():
    for lab in labs:
        LABEL2GROUP[lab] = g

GROUP_IDS = list(GROUP_ID2DESC.keys())
GROUP_ID2IDX = {g: i for i, g in
    enumerate(GROUP_IDS)}
IDX2GROUP_ID = {i: g for g, i in
    GROUP_ID2IDX.items()}
NUM_GROUPS = len(GROUP_IDS)
```

E.3 Label-wise Scoring Prompt

```
Your task is to assign an **independent
    confidence score** in [0, 1]
to EACH fine-grained label in the taxonomy.
    Each score should reflect
how strongly the [Sentence] matches that
    opinion type.

Scoring scale (IMPORTANT):
- 0.00-0.05  : clearly not this opinion type.
- 0.05-0.20  : only very weak / marginal
    evidence.
- 0.20-0.40  : some related evidence, but not
    the main opinion.
- 0.40-0.60  : ambiguous or mixed evidence.
- 0.60-0.80  : clear and good match.
- 0.80-1.00  : very strong and primary match.
```

```
Use the **full range** of scores. Do NOT
    default to 0.0000 for most labels.
For each sentence you should usually:
- assign 1-3 labels with scores >= 0.60,
- assign several other labels with moderate
    scores between 0.10 and 0.60.
Avoid using exactly 0.0000 or 1.0000 unless
    something is logically impossible
or almost certainly true.

Output format (STRICT):
- First line MUST be exactly:
  LABEL_SCORES:
- Then output ONE line for EVERY fine-grained
    label, in the SAME
  order as in the taxonomy (OPINION_LABELS).
- Each of these lines MUST have the form:
  <LABEL_ID>: <SCORE>
  where:
    * <LABEL_ID> is the label id (for example
    A.3-5),
    * there is exactly ONE space after the
    colon,
    * <SCORE> is a decimal number between 0 and
    1 with EXACTLY
      four digits after the decimal point
    (e.g., 0.1375).

Do NOT output any other text, comments,
    explanations, bullets,
markdown, or blank lines.

Now read the instance and output the scores in
    this format only:

{instance_block}
```

E.4 List-wise Scoring Prompt

```
Your task is to assign a **probability score**
    in [0, 1]
to EACH fine-grained label in the taxonomy.
    These scores represent a
**list-wise distribution**: each score reflects
    how strongly the
[Sentence] matches that opinion type **relative
    to all other labels**.

The scores across ALL labels MUST sum to
    approximately **1.0000**
(acceptable range: 0.98-1.02 due to rounding).

Scoring scale (IMPORTANT):
Use the same evidence-strength interpretation
    as independent scoring,
but apply it in a comparative, distributional
    manner:

- 0.00-0.05  : clearly not this opinion type.
- 0.05-0.20  : only very weak / marginal
    evidence.
- 0.20-0.40  : some related evidence, but not
    the main opinion.
- 0.40-0.60  : ambiguous or mixed evidence.
- 0.60-0.80  : clear and good match.
- 0.80-1.00  : very strong and primary match
    (rare in distributional
               scoring, because probabilities
    must sum to 1).

Use the **full range** of probabilities. Do NOT
    default to tiny values
or produce a nearly uniform distribution.

To ensure a meaningful probability distribution:
- Assign **1-3 labels** noticeably higher
    probabilities than all others
  (typically 0.12-0.35, depending on the
    sentence).
- Assign **several labels** moderate
    probabilities (e.g., 0.05-0.15)
  to indicate partial or secondary relevance.
- Assign remaining labels small but non-zero
    probabilities (e.g.,
  0.00-0.05), but avoid making multiple labels
    have identical values.
- The final distribution MUST show **clear
    differentiation** across labels,
  not a flat or near-flat pattern.

Avoid using exactly 0.0000 unless something is
    logically impossible
or almost certainly irrelevant. Avoid
    collapsing all probability mass
into a single label.

Output format (STRICT):
- First line MUST be exactly:
  LABEL_SCORES:
- Then output ONE line for EVERY fine-grained
    label, in the SAME
  order as in the taxonomy (OPINION_LABELS).
- Each of these lines MUST have the form:
  <LABEL_ID>: <SCORE>
  where:
    * <LABEL_ID> is the label id (for example
    A.3-5),
    * there is exactly ONE space after the
    colon,
    * <SCORE> is a decimal number between 0 and
    1 with EXACTLY
      four digits after the decimal point
    (e.g., 0.1375).

Do NOT output any other text, comments,
    explanations, bullets,
markdown, or blank lines.

Now read the instance and output the scores in
    this format only:

{instance_block}
```

E.5 Path-wise Scoring Prompt

```
Your task is to output two kinds of numeric
    scores for the [Sentence]:

1) An independent confidence score for each
    opinion-object group.
2) An independent confidence score for each
    fine-grained opinion label.

Use the SAME scoring scale for BOTH group-level
    and label-level scores.

Scoring scale (IMPORTANT):
- 0.00-0.05  : clearly not this opinion type.
```

```
- 0.05–0.20  : very weak / marginal evidence.
- 0.20–0.40  : some related evidence, but not
    the main opinion.
- 0.40–0.60  : ambiguous or mixed evidence.
- 0.60–0.80  : clear and good match.
- 0.80–1.00  : very strong and primary match.

Use the **full range** of scores. Do NOT
    default to 0.0000 for most groups
or labels.
For each sentence you should usually:
- assign 1–2 groups with scores >= 0.60,
- assign 1–3 labels with scores >= 0.60,
- assign several other groups and labels with
    moderate scores
  between 0.10 and 0.60.
Avoid using exactly 0.0000 or 1.0000 unless a
    group or label is
logically impossible for the sentence, or
    almost certainly the
primary and exclusive match.

Output format (STRICT):
- First line MUST be exactly:
  GROUP_SCORES:
- Then output ONE line for EVERY group id, in
    the SAME order as in
  the taxonomy (keys of GROUP_ID2DESC):
  <GROUP_ID>: <SCORE>
- Then output ONE line:
  LABEL_SCORES:
- Then output ONE line for EVERY fine-grained
    label, in the SAME
  order as in the taxonomy (OPINION_LABELS):
  <LABEL_ID>: <SCORE>

Formatting rules:
- Scores must be decimals in [0, 1] with
    EXACTLY four digits
  after the decimal point (e.g., 0.1375).
- There must be exactly ONE space after each
    colon.
- Do NOT output any other text, comments,
    explanations, bullets,
  markdown, brackets, or blank lines.

Now read the instance and output the group and
    label scores
in this format only:

{instance_block}
```

E.6 Pairwise Refinement Prompt

```
Your task is to make a **pairwise comparative
    judgement** between two
candidate fine-grained opinion labels (Label-A
    and Label-B), each with
its definition.

Judge **which label provides a better
    interpretation** of the opinion
expressed in the [Sentence]. The [Context] may
    assist disambiguation,
but explicit evidence in the [Sentence] has
    higher priority.

Decision rules:
- Treat the task as **relative comparison**,
    not absolute evaluation.
- Focus only on the semantic distinctions
    between A and B.
- Prefer the label that better captures the
    author's intended meaning.
- If the support for both labels is genuinely
    similar  output ~0.50.

Output format (STRICT):
Return exactly ONE line:
  P_A: <NUMBER>
where <NUMBER> is in [0, 1] with EXACTLY four
    decimals.
P_A = the degree to which Label-A is a **better
    fit than Label-B**
in this pairwise comparison.

Calibration guide (relative judgement):

- There is NO default preference for either
    label.
  Label-A is NOT assumed to be better than
    Label-B.

- Use 0.5000 ONLY when the evidence for Label-A
    and Label-B is genuinely
  indistinguishable. This should be rare.

- If the evidence leans toward Label-B, the
    score MUST be below 0.50.
  Do NOT stay above 0.50 out of caution.

- 0.48–0.52 : no clear advantage / almost
    indistinguishable

- 0.52–0.65 : slight preference for A
- 0.65–0.80 : clear preference for A
- >0.80     : strong / dominant preference for
    A (rare)

- 0.35–0.48 : slight preference for B
- 0.20–0.35 : clear preference for B
- <0.20     : strong / dominant preference for
    B (rare)

Prefer moderate values when uncertain; avoid
    extreme values unless the
evidence is overwhelmingly one-sided.

Do NOT output any explanations or extra text.

{instance_block}

[Label-A]
ID: {labA_id}
Definition: {labA_def}

[Label-B]
ID: {labB_id}
Definition: {labB_def}
```