Normalizing section names and structure of scientific articles

Nicolau Duran-Silva[1,2], **César Parra-Rojas**[1],
Julian Moreno-Schneider[3], **Georg Rehm**[3]
[1]SIRIS Lab, Research Division of SIRIS Academic, Barcelona, Spain
[2]LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain
[3] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany
{nicolau.duransilva, cesar.parra}@sirisacademic.com,
{julian.moreno_schneider, georg.rehm}@dfki.de

Abstract

The growing amount of scientific literature has increased the need for automatic methods that can retrieve, process, and exploit scholarly content. In this work, we explore section name normalization and hierarchy prediction for scientific articles using a two-level taxonomy. We compare independent, sequential classification models, and generative large language models on the SASC dataset. Results show that classification approaches, particularly sequential models that employ document-level context, consistently outperform generative methods. Incorporating section content is essential for fine-grained classification, while generative models remain limited in zero-shot settings. Our experiments highlight the importance of structure-aware modelling for large-scale scholarly document processing, and the importance of section normalization for the development of advanced research mapping and research assessment tools.

Keywords: Section classification, LLM, Hierarchical text classification, Scholarly document processing

1. Introduction

The rapid growth of scientific literature (Nane et al., 2023) has intensified the need for automated systems that can effectively process, interpret, and reason over full-text research articles (Ghosal et al., 2025; Frommholz et al., 2025). Modern NLP applications increasingly operate on entire scientific papers rather than isolated sentences or abstracts, requiring representations that preserve discourse structure and semantic context (Wadden et al., 2022; Skarlinski et al., 2024; Duan et al., 2025; Keerthana and Gupta, 2025).

Scientific articles are inherently organized into sections that reflect the logical flow of the research process, such as background, methodology, experimental evidence, and interpretation (Sollaci and Pereira, 2004). Accurately identifying the semantic role of these sections is essential for a wide range of downstream tasks, including section-aware retrieval, scientific question answering, summarization, and knowledge graph construction (Accuosto and Saggion, 2019). However, in practice, many scientific parsing tools produce incomplete or inconsistent structural information. This issue is particularly pronounced for documents with complex layouts, such as two-column formats, dense figures, or non-standard templates, where document parsing tools may fail to correctly recover section boundaries and hierarchy. Approaches that infer section roles directly from textual content, as explored in this work, can help reconstruct document structure in such scenarios and provide a complementation to article parsing. Furthermore, LLMs ingest textual data, which can prefer the ingestion of documents as flat or markdown text (Team, 2024). This is particularly relevant with the emergence of retrieval-augmented generation (RAG) systems and research agents that rely on structured textual inputs, when section boundaries and their hierarchical relationships are missing or ambiguous, these systems risk losing critical contextual signals (Keerthana and Gupta, 2025). The problem is further exacerbated in non-normative papers, where section titles deviate from standard conventions, making cross-document alignment more complex.

Section-level structure plays an important role in the development of research assessment and scientometrics, supporting analysis of citations in context, distinguishing whether a reference supports background, methods, or results. Such applications depend on reliable section normalization and hierarchy reconstruction across heterogeneous scientific documents (Vergoulis et al., 2022).

In this work, we explore the possibility of recovering the structure of scientific articles from text at the section level. Beyond improving document parsing, section normalization enables a range of downstream applications that rely on structured scientific content (e.g. RAG, scientific question answering, or scientometric indicator calculation). We address section normalization as a hierarchical classification task under a two-level taxonomy and compare independent, sequential, and generative modeling paradigms. Our results highlight the importance of document-level context and show that

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 218–224
12 May 2026.

structure-aware discriminative models outperform generative approaches in zero-shot settings.

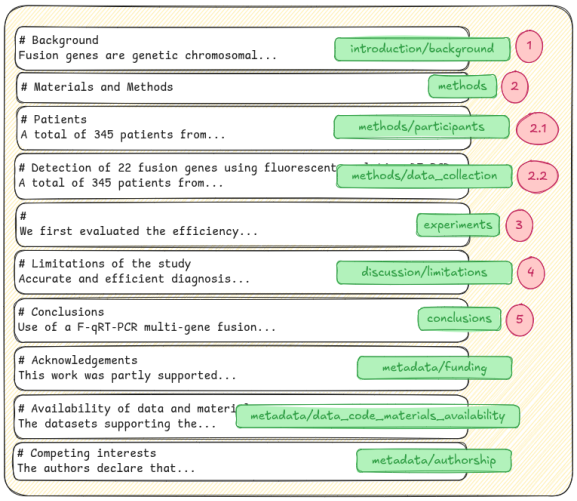

Figure 1: Example of a publication structure annotated with section label and hierarchical level from the SASC Dataset[1].

2. Tasks and Data

We study section-level structure in scientific articles through hierarchical section name normalisation, which aims to map raw section titles (e.g., "Experimental Setup") to a predefined set of canonical section labels. The task is formulated as multi-class classification at two levels of granularity:

- Level-1: coarse roles (e.g., Methods, Experiments);
- Level-2: fine-grained roles (e.g., Data Preprocessing, Evaluation).

Section hierarchy modelling captures the hierarchical organization of sections within a document. However, although the dataset includes explicit section numbering, we restrict our study to a two-level hierarchy (main sections and subsections) based on the 2-level section taxonomy. All deeper levels are collapsed into their immediate parent. Hierarchy is therefore evaluated through level prediction rather than explicit section number prediction.

2.1. Dataset

Experiments are conducted on the Scientific Article Section Classification (SASC) dataset[1] , a large-scale collection of scientific articles segmented by section. The dataset contains 4,896 full-text papers (approximately 117k sections) from different domain splits, including Energy, Cancer, Neuroscience, Transport, and a general science sample.

[1]Available at https://github.com/sirisacademic/sasc

Each section contains raw section title, raw section content, raw explicit section number (when available), semantic labels (level 1 and 2), and hierarchical number. The section taxonomy covers the following Level-1 section categories and their associated Level-2 subcategories are listed below:

- Abstract
- Introduction: background, problem statement, objectives, contributions, outline
- Related Work: literature review, theoretical framework, gaps identified
- Method: study design, participants, data collection, procedures, statistical analysis, ethical considerations, materials and instruments, data preprocessing, evaluation, case study
- Experiment: descriptive statistics, main findings, secondary findings, tables and figures, statistical tests, evaluation
- Discussion: interpretation, comparison with literature, implications, limitations, strengths, future work
- Conclusion: summary, key findings, contributions, final remarks, future directions and recommendations
- Appendix / Miscellaneous: supplementary data, additional analyses, technical details, code and materials, ethics and consent, compliance statement
- Metadata: acknowledgments, authorship, funding, license, data/code availability, publisher note
- References
- Other

3. Methods

We explore section name normalization under different paradigms that differ in how they encode section content and document context.

3.1. Independent classification

We formulate section name normalization as a multi-class classification task, where each section is processed independently. The input to the model consists of the section header concatenated with a truncated snippet of section content separated by a special token:

```
[CLS] sect_header [SEP] sect_content[:50]
```

To control input length and reduce noise, we use only the first 50 tokens of the section content. The resulting sequence is encoded using a language model pretrained on scholarly documents (SciBERT) (Beltagy et al., 2019), and the representation of the `[CLS]` token is passed to a linear classification head. We evaluate three variants: (i) header-only input, (ii) header plus content input, (iii) content

only (filtering only sections with content available). While this approach captures local lexical and semantic elements, it does not exploit dependencies between sections within the same document. The number of the section, for those available, could be a relevant input signal. However, we suspect this could introduce a bias due to the lack of variability and fixed correspondence for some predominant journal structures.

3.2. Sequential classification

To incorporate document-level structure explicitly, we extend the independent classifier into a sequential labelling framework. In this setting, all sections of a document are labelled jointly. Each section is first encoded independently using the same input format as above (header + truncated content) with SciBERT (Beltagy et al., 2019). The resulting section embeddings are then passed to a Transformer encoder (Vaswani et al., 2017) operating over the ordered sequence of sections in a document. This second-stage encoder enables each section representation to attend to other sections, capturing ordering and global patterns. A shared classification head is applied to each contextualized section embedding to predict the normalized label. This architecture mirrors token-level sequence labelling, with sections playing the role of tokens and documents acting as sequences.

3.3. Generative section classification

We additionally explore a generative formulation of section name prediction. In contrast to discriminative classifiers, generative models predict section labels as free-form text conditioned on natural language prompts, which are subsequently normalized to the predefined taxonomy. We explore here two generative settings: *document-level* and *section-level* settings. In the *document-level* setting, all sections of a scientific article are classified in a single model run. The ordered list of section headers, optionally augmented with short content snippets, is provided as input, and the model is prompted to output a sequence of normalized labels in the same order. This formulation enables the model to exploit global document context and implicit structural regularities. In the *section-level* setting, sections are classified sequentially, one at a time, using a local context window. For each section, the model receives the current section header and truncated content, together with the header of the previous and next sections, and—when available—the predicted label of the previous section. This setting approximates incremental document processing to prevent hallucinations in small models while retaining limited contextual information.

In both cases, prompts explicitly enumerate the allowed label taxonomy and enforce output constraints to reduce generation variability. Model outputs are post-processed through a normalization procedure that maps raw generations to valid joint labels (Level-1 and Level-2), with fallback rules applied in case of malformed or out-of-vocabulary predictions. We evaluate two instruction-tuned LLaMA models (Grattafiori et al., 2024), 1B and 8B, under identical prompts, reported in Appendix A, to assess the robustness of generative approaches relative to discriminative baselines. We focus on a zero-shot setting to provide a controlled comparison with discriminative models. However, exploring fine-tuning strategies for generative models is left as future work.

4. Experiments

4.1. Evaluation & Setup

Experiments are conducted on the SASC dataset using document-level train, validation, and test splits. We evaluate section name normalization at two levels of granularity. We additionally report joint correctness (ALL), which requires both Level-1 and Level-2 predictions to be correct. For classification-based models, we report Accuracy and Macro-F1. For generative models, we compute the same section-level metrics after parsing the generated outputs, enabling direct comparison across paradigms. All models are evaluated using identical test sets and label taxonomies.

4.2. Results

Table 1 summarizes the main results. Discriminative models outperform generative approaches across all settings, with sequential models consistently improving over independent classifiers, indicating the benefit of document-level context. Incorporating section content yields substantial gains, particularly for both levels, with content being potentially more relevant than header only. Generative models show lower overall performance under the zero-shot setting, particularly on fine-grained labels, although the LLaMA 8B model performs substantially better than the 1B variant. Especially in the section-level setting, the larger model performs better than the smaller one; however, in the *document-level* setting, the smaller model seems to work better. Unfortunately, these lower perfromance of generative experiments appears to be related to limitations of their model size, such as output processing, taxonomy retention, and hallucinations. Appendix B reports detailed performance by category.

We also analyse performance across scientific domains, reported in Table 2. This indicates that

Model	Paradigm	Input	Accuracy			Macro-F1		
			L1	L2	ALL	L1	L2	ALL
SciBERT	Independent	Header only	0.650	0.491	0.481	0.524	0.339	0.297
SciBERT	Independent	Header + Content	0.777	0.594	0.579	0.672	0.436	0.392
SciBERT	Independent	Content*	0.744	0.521	0.515	0.621	0.338	0.300
SciBERT	Sequential	Header	0.762	0.559	0.547	0.621	0.428	0.396
SciBERT	Sequential	Header + Content	0.797	0.602	0.593	0.683	0.469	0.438
LLaMA-1B	Generative	Header + Content (full)	0.178	0.023	0.089	0.118	0.031	0.036
LLaMA-1B	Generative	Header + Content (section)	0.073	0.001	0.072	0.014	0.001	0.003
LLaMA-8B	Generative	Header + Content (full)	0.460	0.138	0.132	0.311	0.109	0.101
LLaMA-8B	Generative	Header + Content (section)	0.538	0.291	0.251	0.476	0.262	0.226

Table 1: Section Name Normalisation results on the SASC test set. We report Accuracy and Macro-F1 for Level-1 (L1), Level-2 (L2), and joint predictions (ALL). Sequential models exploit document-level context, while generative models perform list-to-list normalization. From *Content-only* we remove all those sections without content, assuming that results are not fully comparable because some challenging cases are removed.

Domain	Accuracy			Macro-F1		
	L1	L2	ALL	L1	L2	ALL
cancer	0.918	0.739	0.715	0.509	0.708	0.485
energy	0.862	0.681	0.680	0.532	0.668	0.499
neuroscience	0.862	0.681	0.680	0.532	0.668	0.499
transport	0.668	0.615	0.450	0.414	0.426	0.390
general	0.693	0.648	0.515	0.443	0.501	0.409

Table 2: Domain-wise performance of the sequential SciBERT model with header and content input.

domains with more standardized article structures and section structure and naming conventions (e.g., cancer and neuroscience) achieve higher accuracy and F1 scores, while domains with greater variability in paper structure (e.g., general domain and transport) are more challenging.

5. Discussion

This work highlights that predicting the semantic structure of scientific articles remains a challenging and unresolved task, even within relatively close scientific domains. While the datasets considered related disciplines, substantial variability in section naming and organization persists, particularly at finer levels of granularity. As a result, current models struggle to generalize consistently across domains with less standardized writing conventions.

Our results show that Level-1 classification is considerably easier than Level-2, reflecting the higher level of abstraction and semantic overlap among fine-grained sections. Models relying only on section headers as input are prone to lexical bias, often learning canonical titles without sufficient contextual grounding. Incorporating section content substantially improves performance, confirming that semantic information beyond headings is critical for reliable classification, as well as the use of signals from document structure.

Generative models exhibit a clear performance gap between coarse- and fine-grained predictions, suggesting limitations in handling with a large number of labels in a zero/few-shot manner. While much larger models could mitigate this issue, their computational cost for processing millions of articles raises practical concerns. This could indicate that lightweight fine-tuning of smaller generative models could bring considerable gains, which will be included as future work.

Overall, this study provides a comparison of architectural paradigms for two-level section classification. Future work includes exploring transfer learning across disciplines, improving efficiency for large-scale processing, and balancing performance gains against computational and environmental costs when deploying larger language models.

5.1. Error analysis

As part of the error analysis, we report in Figure 2 the level-1 confusion matrix for the independent classification model using section headers and content. The results show that most errors arise from confusions with the “Other” and “Abstract” categories, and that sections are frequently misclassified into adjacent categories in the taxonomy, such as “Related Work” vs. “Introduction” or “Method”, and “Discussion” vs. “Experiment”. This pattern suggests that improving the separation between neighbouring categories in the taxonomy during training could represent a promising direction for improvement.

While sequential models may be sensitive to error propagation from earlier sections, our results indicate that the benefits of incorporating document-level context outweigh this limitation. A more detailed analysis of error propagation is left for future

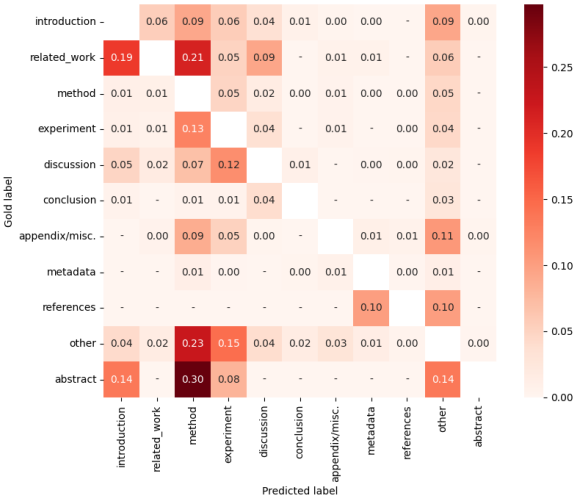

Figure 2: Confusion matrix at Level 1 for the independent classifier based on header + content.

work.

6. Conclusions

In this work, we address section name normalization and hierarchy prediction for scientific articles, comparing independent, sequential, and generative modelling paradigms under a two-level taxonomy. Our experiments show that discriminative approaches based on SciBERT outperform generative models, with sequential document-level modelling providing clear benefits by exploiting structural information across sections. Incorporating section content improves performance, while header-only representations suffer from lexical ambiguity and overlap between adjacent categories. Generative models, despite their flexibility, remain limited in zero-shot settings, particularly for detailed labels, and are sensitive to hallucinations and output normalization. However, differences across scientific domains remain a challenge. Overall, these findings highlight the importance of structure-aware discriminative models for large-scale scholarly document processing and point to future directions in improving taxonomy separation, domain transfer, and computational efficiency.

7. Acknowledgements

Supported by the Industrial Doctorates Plan of the Department of Research and Universities of the Generalitat de Catalunya, by Departament de Recerca i Universitats de la Generalitat de Catalunya (grant reference 2022/DI /00017). This work was co-funded by the EU HORIZON project SciLake (Grant Agreement 101058573) and the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS)[2] as part of the non-profit association National Research Data Infrastructure (NFDI e. V.). The consortium is funded by the Federal Republic of Germany and its states through the German Research Foundation (DFG) project NFDI4DS (no. 460234259). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funders. Neither the European Union nor the granting authority can be held responsible for them.

We thank the anonymous reviewers for their constructive feedback and suggestions, which improved the clarity and quality of this work.

8. Bibliographical References

Pablo Accuosto and Horacio Saggion. 2019. Discourse-driven argument mining in scientific abstracts. In *International Conference on Applications of Natural Language to Information Systems*, pages 182–194. Springer.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Decheng Duan, Jitong Peng, Yingyi Zhang, and Chengzhi Zhang. 2025. SciNLP: A domain-specific benchmark for full-text scientific entity and relation extraction in NLP. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14473–14486, Suzhou, China. Association for Computational Linguistics.

Ingo Frommholz, Philipp Mayr, Guillaume Cabanac, Suzan Verberne, and Christin Katharina Kreutz. 2025. The first workshop on scholarly information access (scolia). In *European Conference on Information Retrieval*, pages 326–331. Springer.

Tirthankar Ghosal, Philipp Mayr, Anita De Waard, Aakanksha Naik, Amanpreet Singh, Dayne Freitag, Georg Rehm, Sonja Schimmler, and Dan Li. 2025. Overview of the fifth workshop on scholarly document processing. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 1–6.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur,

[2] https://www.nfdi4datascience.de

Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Garapati Keerthana and Manik Gupta. 2025. Clirag: A retrieval-augmented framework for clinically structured and context aware text generation with llms. *arXiv preprint arXiv:2507.06715*.

Gabriela F Nane, Nicolas Robinson-Garcia, François van Schalkwyk, and Daniel Torres-Salinas. 2023. Covid-19 and the scientific publishing system: growth, open access and scientific fields. *Scientometrics*, 128(1):345–362.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*.

Luciana B Sollaci and Mauricio G Pereira. 2004. The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey. *Journal of the medical library association*, 92(3):364.

Deep Search Team. 2024. Docling technical report. Technical report.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thanasis Vergoulis, Serafeim Chatzopoulos, Kleanthis Vichos, Ilias Kanellos, Andrea Mannocci, Natalia Manola, and Paolo Manghi. 2022. Bip! scholar: a service to facilitate fair researcher assessment. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Appendix A. Prompt Templates for Generative Section Classification

We employ structured prompt templates to guide large language models in performing section name normalization under explicit output constraints. Prompts enumerate the allowed section taxonomy and enforce deterministic, schema-conform outputs.

Document-Level Section Classification Prompt

```
You are an expert scientific document
     section normalizer.

TASK:
For each section below, assign exactly ONE
     label from the ALLOWED LABELS list.

ALLOWED LABELS:
{L1 | L2 taxonomy listing}

RULES:
- Use "L1|L2" when a specific subtype
     applies.
- Use only "L1" when no subtype is
     appropriate.
- Output exactly N labels, one per section,
     in order.
- Output ONLY a JSON array of strings. No
     explanations.

DOCUMENT:
[1] HEADER: {SECTION_HEADER_1} | CONTENT: {
     CONTENT_SNIPPET_1}
[2] HEADER: {SECTION_HEADER_2} | CONTENT: {
     CONTENT_SNIPPET_2}
...
[N] HEADER: {SECTION_HEADER_N} | CONTENT: {
     CONTENT_SNIPPET_N}

OUTPUT:
["label_1", "label_2", ..., "label_N"]
```

Section-Level Prompt with Local Context

```
You are an expert scientific document
     section normalizer.

TASK:
Classify the CURRENT section into exactly
     ONE label from the ALLOWED LABELS.

ALLOWED LABELS:
{L1 | L2 taxonomy listing}

SURROUNDING CONTEXT:
PREVIOUS SECTION: {PREV_HEADER} (label: {
     PREV_LABEL})
NEXT SECTION: {NEXT_HEADER}

CURRENT SECTION:
HEADER: {CURRENT_HEADER}
CONTENT: {CONTENT_SNIPPET}

OUTPUT:
Return ONLY a single label string (e.g., "
     method|data_collection",
"introduction|background", or "references").
```

9. Appendix B. Performance by category

Table 3 reports per-category performance, showing higher accuracy for frequent and structurally well-defined sections (e.g., Method, Conclusion, Metadata) and lower performance for heterogeneous

categories.

Section	P	R	F1
Introduction	0.80	0.65	0.72
– Background	0.97	0.68	0.80
– Contributions	1.00	0.05	0.09
– Objectives	0.49	0.30	0.38
– Outline	0.00	0.00	0.00
– Problem Statement	0.37	0.23	0.29
Related Work	0.31	0.39	0.34
– Gaps Identified	0.00	0.00	0.00
– Literature Review	0.80	0.38	0.51
– Theoretical Framework	0.50	0.14	0.21
Method	0.80	0.86	0.83
– Case Study	0.91	0.68	0.78
– Data Collection	0.59	0.55	0.57
– Data Preprocessing	0.59	0.49	0.53
– Ethical Considerations	0.95	0.50	0.66
– Evaluation	0.51	0.39	0.44
– Materials And Instruments	0.57	0.50	0.53
– Participants	0.82	0.78	0.80
– Procedures	0.75	0.63	0.68
– Statistical Analysis	0.90	0.74	0.81
– Study Design	0.64	0.38	0.48
Experiment	0.79	0.77	0.78
– Descriptive Statistics	0.47	0.43	0.45
– Evaluation	0.38	0.37	0.38
– Main Findings	0.68	0.49	0.57
– Secondary Findings	0.47	0.32	0.39
– Statistical Tests	0.64	0.18	0.28
– Tables Figures	0.78	0.45	0.57
Discussion	0.70	0.71	0.71
– Comparison Literature	0.49	0.29	0.36
– Future Work	0.58	0.38	0.46
– Implications	0.58	0.26	0.36
– Interpretation	0.90	0.63	0.74
– Limitations	0.92	0.85	0.89
– Strengths	0.00	0.00	0.00
Conclusion	0.91	0.92	0.91
– Contributions	0.00	0.00	0.00
– Final Remarks	0.89	0.39	0.54
– Future Directions Recommendations	0.67	0.62	0.65
– Key Findings	1.00	0.14	0.25
– Summary	0.85	0.90	0.87
Appendix/Misc.	0.86	0.73	0.79
– Additional Analyses	0.00	0.00	0.00
– Code And Materials	1.00	0.43	0.60
– Compliance Statement	1.00	0.25	0.40
– Ethics And Consent	1.00	0.93	0.96
– Supplementary Data	0.92	0.71	0.80
– Technical Details	0.61	0.23	0.34
Metadata	0.98	0.96	0.97
– Acknowledgments	1.00	0.99	0.99
– Authorship	1.00	0.98	0.99
– Data Code Materials Availability	1.00	0.98	0.99
– Funding	1.00	0.94	0.97
– License	0.33	0.17	0.22
– Publisher Note	0.99	0.85	0.92
References	0.35	0.80	0.48
Other	0.44	0.46	0.45
Abstract	0.68	0.35	0.46

Table 3: Per-category precision (P), recall (R) and F1-score, using the independent classification model with header + content, reporting results at level 1 and 2.