Benchmarking LLMs for ARR Area Assignment: Evidence and Implications for Assignment Strategies

Eileen Bingert, Diego Alves, Stefania Degaetano-Ortlieb
Saarland University, Department of Language Science and Technology
Campus A2.2, 66123 Saarbruecken
s8eibing@teams.uni-saarland.de, diego.alves@uni-saarland.de, s.degaetano@mx.uni-saarland.de

Abstract

We study how large language models (LLMs) perform at assigning review areas, specifically for the Association of Computational Linguistics (ACL) Rolling Reviews (ARR), using titles and abstracts from 2020–2025. We compare multiple LLMs and prompting schemes (zero/few-shot; with/without ARR keywords; each-category variants) and analyze per-area scores, error overlap, and confusion matrices. One-shot prompting (with OpenAI-gpt-oss-20b) tends to perform best, while injecting ARR keywords often lowers accuracy. Task-bounded areas (e.g., MT, IE, QA, Summarization) are predicted more reliably, whereas broad, cross-cutting labels (e.g., Resources and Evaluation, NLP Applications) are frequently conflated, indicating taxonomy ambiguity rather than solely model limitations. We recommend hierarchical or primary-plus-secondary labels to work towards reducing ambiguity and improving reviewer matching in the future. Our dataset, methods, and findings offer a reproducible baseline for area selection support in conference workflows such as for ACL.

Keywords: LLM benchmarking, Zero-/few-shot prompting evaluation, ACL area classification

1. Introduction

A known problem in the academic community for reviewer assignment is to generate a system of areas to match reviewer expertise. As this is not a trivial task, given that the choice directly affects reviewer assignment to papers, keywords have been established to help authors select the right area for their papers. In this paper, we run a pilot test to assess whether the application of large language models (LLMs) might indicate insights on area assignment and its development over time. For this, we run a case study on papers from the Association for Computational Linguistics (ACL), a long-running, richly studied scholarly ecosystem. Quite some endeavours have been going into updating keywords, areas, and tracks in order to match the ever changing landscape of ACL[1] and related conferences[2] (Rogers et al., 2023). Since 2021, the ACL Rolling Review (ARR) introduced a common, evolving *area taxonomy* with public area descriptions and keywords to standardize reviewer matching across ACL-family conferences. Statistics (see Rogers et al. (2023) Table 1) show that while some areas are relatively specific and rather small, areas with less topical focus such as *Resources and Evaluation* or *NLP Applications* act as collective bins for various topics, making reviewer assignment more problematic. Thus, area labels can be *coarse* and partially overlapping, creating ambiguity for authors and area chairs. Also, area choice interacts with automated affinity matching (such as the Toronto Paper Matching System (TPMS)) and chair policies, so misalignment between an article's content and its declared area can influence both the reviewer pool and evaluation outcomes. Moreover, human-curated keywords may incompletely capture area semantics and can again overlap across areas (see arr (2025b,a); emn (2025), e.g., *evaluation* as a keyword across areas), potentially biasing both manual choice and automated matching.

In this paper, we ask whether given paper titles and abstracts, LLMs would classify papers to the chosen ACL areas. Prompted LLMs could provide a complementary signal: given the same title and abstract string available at submission time, they operationalize a content-driven mapping into the ARR taxonomy. By benchmarking different prompting schemes against author-declared areas, we can (i) estimate attainable agreement from content alone, and (ii) identify systematic confusions that may indicate where areas are semantically conflated or keywords mis-specify scope. For this, we use the main conference titles and abstracts from the ACL, EACL and NAACL venues from 2020 up to 2025 with 558 papers in total. We compare the results of different LLMs given also differing prompting strategies (e.g., with and without keywords). Results show how general areas, especially, NLP applications, encompasses papers, which would well fit other areas, as well as how areas might actually be closely related and could be conflated.

[1] https://aclrollingreview.org/areas
[2] https://2025.emnlp.org/track-changes/

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 13–24
12 May 2026.

2. Theoretical Background

Our study connects to two strands of prior work: (i) the use of LLMs to support topic selection and document triage in scientific workflows and (ii) prompt-based control of LLMs for classification.

2.1. LLMs for Topic Selection and Scholarly Triage

LLMs are increasingly used as high-recall screeners and labelers over titles and abstracts in systematic reviews and literature audits, offering substantial efficiency gains while maintaining competitive recall when prompts and consensus protocols are carefully designed (Dennstädt et al., 2024; Joos et al., 2024). Beyond screening, LLMs have been shown to support structured topic and label assignment: Kuzman and Ljubešić (2025) propose a teacher–student framework for news topic classification in under-resourced languages, demonstrating that LLM-generated labels can effectively train downstream classifiers within the IPTC hierarchical taxonomy (e.g., politics, economy, society). Similarly, Zhu et al. (2025) use LLM-derived representations for hierarchical taxonomy induction and scientific paper clustering, achieving state-of-the-art taxonomy coherence and interpretability. Together, these findings support the feasibility of LLM-driven constrained topic assignment in structured label spaces.

In the area of NLP, Ahmad et al. (2024a) created a hierarchical taxonomy and manually annotated 1500 ACL Anthology publications with their main contributions using CL/NLP topics and sub-topics. They then compare two different approaches for classifying publications with a fine-tuned SciNCL model (Ahmad et al., 2024b). The weakly supervised X-transformer model outperforms the competing approach as well as the baseline, reaching a micro precision of 0.4391. This suggests that models are able to correctly assign labels in an extreme multi-label classification task focusing on the area of NLP and the ACL anthology while also highlighting the difficulty of such a task with limited training data.

In scenarios with no or very limited data for fine-tuning, generative LLMs offer an accessible alternative for label generation and classification. Kuzman et al. (2023) report promising results using GPT-3.5 for genre identification, while Säily et al. (2025) evaluate newer GPT models for generating genre labels for novels in the Corpus of Historical American English. Focusing on scientific texts in contemporary English and German, the second LLMs4Subjects shared task (D'souza et al., 2025) included a subtask on automatic domain identification, requiring systems to assign one or more labels from 28 predefined categories (Ho, 2025; Shirali et al., 2025).

2.2. Prompting Strategies

For improving the output of LLMs, prompt engineering has proven to be an effective tool. The modification of the input prompt is a relatively simple way in comparison to other techniques used to increase the performance as it does not require the retraining or fine-tuning of models. Therefore, it is often accessible to users without coding abilities as the prompts are written in natural language (Sahoo et al., 2024). Thus, prompt engineering has become a primary lever for steering LLM behavior without parameter updates, with established families including zero-shot, few-shot, instruction-style prompting, and variants that manipulate format, label space exposure, and task decomposition (Schulhoff et al., 2024; Sahoo et al., 2024). A prompt template typically consists of some context or input and the desired output. This template is then repeated n-times with $n>0$ for few-shot prompting and $n=0$ for zero-shot prompting (Brown et al., 2020). Zero-shot prompting solely contains a description of the task with no additional exemplary input data. Therefore, the LLM relies only on pre-existing knowledge and understanding to perform the required action (Sahoo et al., 2024). Few-shot prompting operationalizes *in-context learning* (ICL), where models condition on a small set of input–output exemplars (also called demonstrations) to induce a task-specific mapping at inference time (Brown et al., 2020). Subsequent work has refined our understanding of why demonstrations help: accurate labels are not always necessary, while exposure to the label inventory, input distribution, and output format often accounts for much of the ICL gains (Min et al., 2022). These findings motivate our controlled variants that (a) vary the number and coverage of demonstrations across ARR areas and (b) expose or hide area keywords to test whether lexical cues aid or distract models during area selection.

One disadvantage of few-shot prompting is the potential inclusion of biases in the input prompt, which then are reflected in the output of the LLM. Therefore, the wording of the input can have significant influence on the model's behaviour. Another is the lengthening of the input prompt that can occur with multiple examples, which then can exceed the input token limit or increase the runtime (Sahoo et al., 2024). We try to account for this by considering varying numbers of examples provided in the prompt, which clearly affect the length of the prompt.

For taxonomy-based classification, two design choices matter. First, demonstrations should represent the *label space* broadly enough to reduce class-imbalance artifacts in ICL (Min et al.,

2022). Second, prompt content can introduce *shortcut features* (e.g., area-specific keywords) that improve surface matching but may reduce robustness when areas overlap semantically; our "with/without keywords" comparison explicitly probes this trade-off (Schulhoff et al., 2024; Sahoo et al., 2024).

3. Data

3.1. Data Compilation

To conduct our experiments, we compile a corpus in English of the relevant data of the ACL, EACL, NAACL main conferences from 2020 to 2025 using the ACL Anthology[3].

	2023	2024	2025	Total
ACL	29	289	5	323
EACL	3	49	0	52
NAACL	0	180	3	183
Total	32	518	8	558

Table 1: Number of papers in the corpus per year and conference that were successfully matched to an ARR area.

In a first step, we download the proceedings. These HTML files encompass all papers from the proceedings for a single event. We only focus on the main conference given that it is the one operating strictly on the ARR area classification for reviewer assignment. We use a Python script to extract the relevant papers (excluding e.g. workshop proceedings).

As the goal of our research is to measure the performance of different LLMs in assigning the different ARR areas, we require a second file with the officially assigned areas as a baseline for comparison. Therefore, in a second step, we download the data of the category *Anonymous Pre-prints* from Open Review Rolling Reviews[4] for the years from 2021 to 2025. For 2020, there is no data available. The areas are chosen by the authors during the submission process. We use a second Python script that compares the titles of the data downloaded in the first step to the titles in these files. We repeat this step for three years for each event/year combination: The year of the conference, as well as the two years before that. For instance, the data for ACL 2024 is compared to the Open Review Rolling Reviews data for 2022, 2023 and 2024. If a match occurs, the area is added to the other metadata (See Table 2) and saved in a separate file. In total, our data set contains 558 abstracts successfully matched to an area[5]

Due to incomplete availability of OpenReview metadata and imperfect title matching, only 558 papers (out of 8,905 papers published, 6%) could be successfully aligned with ARR area labels. This constitutes a limitation of our dataset and likely under-represents the full set of conference papers. Consequently, our study should be viewed as a preliminary investigation into the feasibility and performance of LLM-based ARR area assignment, rather than a definitive large-scale evaluation.

Table 1 reports the number of papers successfully matched to an ARR area per conference and year. As shown, there is a large discrepancy in the number of papers across years: ACL 2024 accounts for 518 papers, while 2023 and 2025 only have 32 and 8 papers, respectively. This difference is primarily due to the inhomogeneous availability of ARR metadata in OpenReview for different years.

3.2. ARR Areas

As stated above, we successfully assigned an ARR area[6] to 558 papers from our data set. The first list of areas and keywords was published in 2023 and then updated in 2024 and 2025. Thus, our data includes 23 areas from different versions, which we combine into a single list. Additionally, we also merge two areas, the two ACL '23 areas *Semantics: Lexical* and *Semantics: Sentence-level Semantics, Textual Inference, and Other Areas* under the new label *Semantics*, as the 2025 version contains a single area with this thematic focus. The keywords were selected by the senior area chairs of each area, based on their knowledge of the area. The ARR states that the list is not exhaustive, but an overview of the most common themes.

Two of the 25 ACL areas we chose to include, namely *Human-centered NLP* and *Language Modeling*, are not included in our data set, likely because they were only present in 2024. We still chose to include the areas in nine of our ten prompting strategies as one minor goal of the evaluation task is to test the accuracy of the tag chosen by the respective paper's authors. Papers that are assigned the area *Special Theme Track* are excluded from the data, as this label is not transparent and changes for each conference. Therefore, a thematic classification is not possible and no keywords are available.

The areas vary in size, the smallest areas *Syntax: Tagging, Chunking and Parsing* and *Phonology, Morphology and Word Segmentation* consist-

[3] https://aclanthology.org/

[4] https://openreview.net/group?id=aclweb.org/ACL/ARR

[5] https://github.com/eilebin/acl-task-data

[6] https://aclrollingreview.org/areas

Category	Example
ID	Borenstein_Arora_Kaffee_Augenstein_NAACL_2025
Title	Investigating Human Values in Online Communities
Authors	Nadav Borenstein, Arnav Arora, Lucie-Aimée Kaffee, Isabelle Augenstein
Year	2025
Event	naacl
Type	long
Abstract	Studying human values is instrumental for cross-cultural research [...]
Area	Computational Social Science and Cultural Analytics

Table 2: Example of the extracted data for each paper. The ID is in a BibTeX-like format, consisting of a maximum of five authors, the event, and year. Its purpose is the identification of the paper.

ing of 3 papers, whereas the biggest area *Resources and Evaluation* encompasses 68 papers. Figure 1 displays the number of papers in each area.

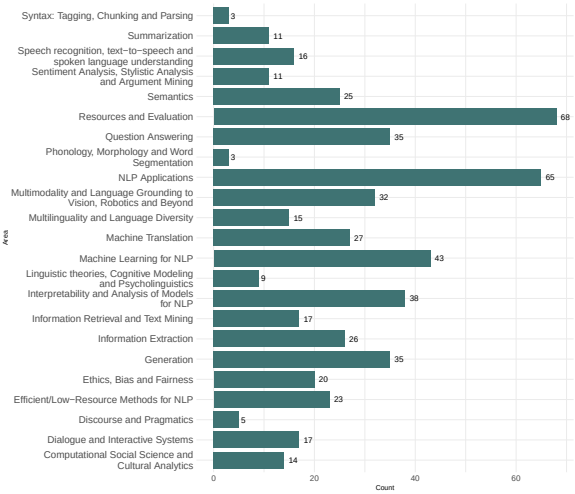

Figure 1: Number of papers across the different ARR areas in the data set.

4. Methodology

4.1. Models

We evaluate the performance of three different models on our dataset. Given that our data set is limited in size, we chose to focus on existing models rather than train an encoder model.

Hermes 2 Pro - Llama-3 8B Hermes 2 Pro - Llama-3 8B[7] (Teknium et al., 2024) was released as an upgrade of Nous Hermes 2, based on Llama-3 (8 billion parameters).

Qwen3-32B Qwen3-32B (Qwen Team, 2025). The model[8] is a 32.8B-parameter causal language model with 64 layers and grouped-query attention (64 query heads and 8 key-value heads). It offers both a thinking and a non-thinking mode, which we indicate by the addition of _Think.

OpenAI-gpt-oss-20b OpenAI-gpt-oss-20b[9] (OpenAI, 2025) is an autoregressive Mixture-of-Experts transformer with 24 layers and 20.91B parameters, released on 5 August 2025. OpenAI-gpt-oss-20b was trained on a text-based data set, focusing on science, code and general world knowledge.

All models were run locally using the Hugging Face Transformers library with automatic GPU device placement. Inference was conducted in bfloat16 precision without additional quantization. Prompts were formatted using the models' native chat templates, and deterministic decoding was used to ensure reproducibility. No sampling-based hyperparameters were added. The generated outputs were saved as YAML files without further post-processing.

4.2. Prompting Strategies

As stated in Section 2, we test different prompting techniques on the models, both with and without human-curated keywords. In all cases, the prompt includes a description of the task and the structure of the input data.

- **zero-shot:** For this, we offer the model an example output with no additional material.
  Example Output (in YAML format):
  article_id:
  acl_area:
- **one shot, three shot, five shot**: This prompt encompasses one (three, five) randomly chosen examples in an input and the required output format. An example consists of the data in

[7] https://huggingface.co/NousResearch/Hermes-2-Pro-Llama-3-8B

[8] https://huggingface.co/Qwen/Qwen3-32B

[9] https://huggingface.co/openai/gpt-oss-20b

Table 2 as the Example Input, as well as the article_id and the area in the Example Output.

- **each category:** In this prompt, we include one example for each category that is included in our curated list of ARR areas.
- **each category_v2:** Here, we only include the ARR areas that are represented in the data in the list of possible areas the models can choose from. Therefore, the number of areas is reduced from 25 to 23, as the areas *Human-centered NLP* and *Language Modeling* are not included in our data set.
- **each category_three:** In this prompt, we include three examples for each area in the data. However, due to the necessity to reduce the length of the prompt, we only give the example output.
- **keywords_zero shot, keywords_one shot, keywords_three shot, keywords_five shot:** In addition to the previous information, the keywords provided by the ARR are added to the prompt.

In Appendix A, we provide the prompt used in the one-shot experiment to illustrate our prompting strategy.

5. Evaluation of Different Prompting Strategies

5.1. Zero-Shot vs. Few-Shot Prompting

As Figure 2 portrays, the accuracy of 2 of the 4 models, i.e. Hermes2Pro and the non-thinking version of Qwen3-32B, increases with a higher number of examples in the prompt.

Considering Table 3, Hermes2Pro gradually increases its performance with the inclusion of more examples into the prompt. Initially, at zero-shot prompting, it classifies 174 papers correctly. It then peaks at five-shot with 196 papers, a 12.64 % increase. The non-thinking version of Qwen3-32B mirrors this behaviour. At zero-shot, the label for 224 papers is correctly selected. This value expands by 14.73 % to 257 at five-shot.

However, for OpenAI-gpt-oss-20b offering more examples does not improve the performance of the LLM. This model peaks at one-shot prompting with 309 correctly classified papers, the highest value in our data, and then declines as more examples are added. At five-shot prompting, 299 areas could be correctly assigned to the papers, a 3.24 % decrease. Similarly, Qwen3-32B_Think reaches its best output performance with only one example in the prompt at 284 papers.

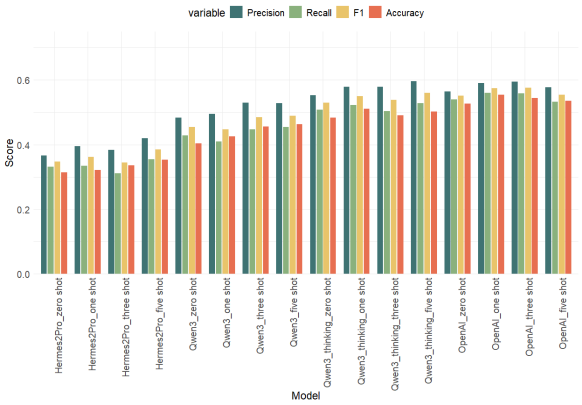

Figure 2: Performance of the four models in terms of macro-averaged precision, recall, F1-score, and accuracy across different prompting strategies without keywords.

The highest recall (0.560) is achieved by OpenAI-gpt-oss-20b at one-shot prompting, reflecting the high accuracy of this specific prompting and model combination. In contrast, Hermes2Pro at three-shot has the lowest recall with 0.312. Notably, the model and few-shot combination with the highest accuracy does not reach the highest macro-averaged precision. Qwen3-32B_Think reaches a precision of 0.597 at five-shot, whereas the best performing OpenAI-gpt-oss-20b achieves a precision of 0.5899 at one-shot.

5.2. All Area Coverage

For the previous experimental setup, we only offered a fixed number of examples in the prompt in total, so not every area was represented. In this step, we increase the number of examples significantly to 23 and 69 respectively, covering all of the areas in the data set.

Considering again Table 3, three of the four models do not display an increase in correctly assigned areas for the each-category prompt. Only the non-thinking version of Qwen3-32B performs better than at zero-shot and few-shot with 271 papers. Compared to five-shot, the best performing experimental set-up for our initial testing, the model increases its correct output by 5.45 %.

Additionally, the exclusion of the two supplementary areas (*Human-centered NLP* and *Language Modeling*) has a positive impact on the performance of all models. The non-thinking version of the Qwen3 model again outperforms its previous results with 276 papers.

As with the previous set-up, OpenAI-gpt-oss-20b does not achieve a rise in correct answers as more examples are added to the prompt. For each category_three, the model even reaches the second-lowest result in total, with 293 papers. This

	Hermes2Pro	Qwen3	Qwen3_Think	OpenAI
		Prompts without keywords		
zero-shot	174 (31.18 %)	224 (40.14 %)	269 (48.21 %)	294 (52.69 %)
one-shot	178 (31.90 %)	236 (42.29 %)	284 (50.90 %)	**309 (55.38 %)**
three-shot	186 (33.33 %)	253 (45.34 %)	273 (48.92 %)	304 (54.45 %)
five-shot	196 (35.13 %)	257 (46.06 %)	279 (50.00 %)	299 (53.58 %)
each category	186 (33.33 %)	271 (48.57 %)	277 (49.64 %)	291 (52.15 %)
each category_v2	193 (34.59 %)	276 (49.46 %)	280 (50.18 %)	300 (53.76 %)
each category_three	226 (40.50 %)	293 (52.51 %)	**306 (54.84 %)**	293 (52.51 %)
		Prompts with keywords		
keywords_zero-shot	205 (36.74 %)	219 (39.25 %)	239 (42.83 %)	238 (42.65 %)
keywords_one-shot	169 (30.29 %)	223 (39.96 %)	251 (44.98 %)	265 (47.49 %)
keywords_three-shot	187 (33.51 %)	230 (41.22 %)	260 (46.59 %)	271 (48.57 %)
keywords_five-shot	196 (35.13 %)	258 (46.24 %)	283 (50.72 %)	**292 (52.33 %)**

Table 3: Model Performance Comparison

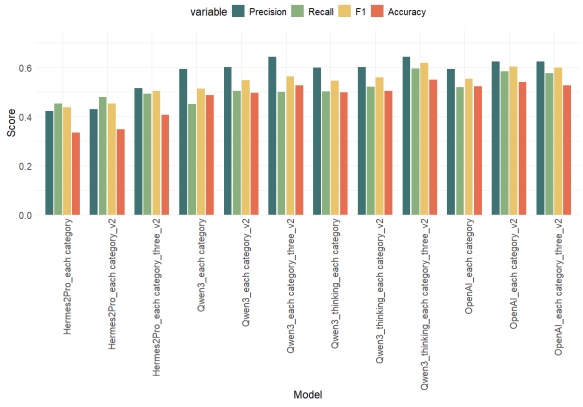

Figure 3: Performance of the four models in terms of macro-averaged precision, recall, F1-score, and accuracy across different prompting strategies with coverage of all areas (each category)

drop is again contrary to the behaviour of the other models that show a positive correlation between the number of examples in the prompt and the performance. Qwen3_Think correctly classifies 306 papers, a 7.75% rise compared to one-shot. Hermes 2 Pro - Llama-3 8B reaches a surge of 15.31% from 196 at five-shot to 226.

In total, the reduction of class-imbalance in the prompt seems to only aid the models if multiple examples for each area are added, as seen in Figure 3. The data for three out of four LLMs suggests that ICL improves with a higher number of examples.

5.3. Prompts with keywords

The inclusion of keywords into the prompt decreases the performance for three of the four models for zero shot, one shot and three shot prompting (see again Table 3). For all of the four variants of shots, the biggest drop can be seen for OpenAI-gpt-oss-20-b model with a decrease of 19.05% for zero shot, 14.24% for one shot, 10.86% for three shot and 2.34% for five shot. Only the Hermes2Pro model reaches better results in two cases, for zero shot and three shot. Additionally, both versions of the Qwen3 model display a better performance for the five shot prompts.

When adding keywords, as presented in Figure 4, the Qwen3 and OpenAI models increase their performance with more examples, a behaviour which is contrary to the results of the experimental setup without keywords for OpenAI-gpt-oss-20b and Qwen3-32B_Think (see Section 5.1). For instance, Qwen3-32B_Think correctly selects the areas of 239 papers for zero-shot prompting. For five-shot its accuracy rises by 18.41% to 283 papers, with no drop for one-shot and three-shot.

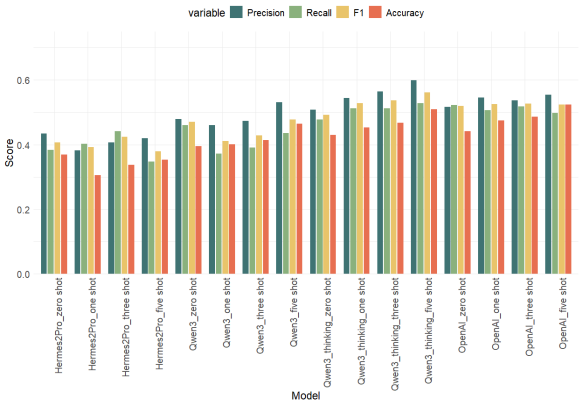

Figure 4: Performance of the four models in terms of macro-averaged precision, recall, F1-score, and accuracy across different prompting strategies with human-curated keywords.

Despite the lower performance in general,

OpenAI-gpt-oss-20b outperforms the other models for one shot, three shot and five shot. With 292 out of 558 (52.33%) correctly classified papers, it reaches the best results for this experimental setup at five shot prompting. For zero-shot, the Qwen 3-Thinking model displays a better performance, with a one paper difference.

In total, the inclusion of keywords seems to not aid the models in the area selection task. This could be due to the human-curated nature of the keywords which might not accurately and fully represent the respective areas. Some keywords are included in multiple areas. For instance, the keyword ”human-in-the-loop” is used for the areas *Dialogue and Interactive Systems* and *Human-Centered NLP*. Additionally, for the area *Linguistic Theories, Cognitive Modeling, and Psycholinguistics* the only keywords are the three parts of the area name.

When comparing our results with previous work, we observe that state-of-the-art performance varies considerably depending on the taxonomy and language. Results reported for the LLMs4Subjects shared task (Ho, 2025; Shirali et al., 2025) are relatively lower than those obtained for our specific task, while Kuzman et al. (2023) report F1 scores above 70 for under-resourced languages. Our task is particularly challenging, as it involves identifying fine-grained research areas within a single scientific field, where substantial inter-topic overlap is likely. Furthermore, the labels are derived from authors' self-assessments, which may introduce subjective bias and additional noise into the training and evaluation data.

6. Evaluation of Results per Area

To assess the performance differences between the areas, we now focus on a micro-level analysis of the results for OpenAI-gpt-oss-20b at one-shot prompting, the best performing model and prompting technique combination in our data (see Figure 5). Additionally, we examine a confusion matrix of the results to trace the discrepancy between declared and assigned areas (see Figure 6).

Figure 5 illustrates the results of the selection task divided by area. Notably, for two areas, *Computational Social Science and Cultural Analytics* and *Phonology, Morphology and Word Segmentation*, the model reaches a perfect precision score. However, for the latter category, this result is only based on one successfully assigned paper. For both areas, this is not reflected in the recall score. The lowest precision scores are noted for the areas *Syntax: Tagging, Chunking and Parsing* and *Semantics*. Similarly, the recall for *Semantics* is also the lowest, along with *NLP Application*. As Figure 6 shows, a higher number of papers from *NLP Applications* were marked as *Generation* instead of the correct area. Seven papers from *Semantics* were categorized as *Machine Learning for NLP*, while four were given the correct area. The highest recall is reached for *Machine Translation* and *Summarization*. *Machine Translation* also has the highest F1-score.

The confusion matrix gives us a more detailed breakdown. Task-bounded areas show a crisp diagonal (e.g., machine translation, information extraction, question answering, summarization), indicating balanced precision and recall. By contrast, broad or cross-cutting areas behave as catch-alls and create asymmetric errors. *Resources and Evaluation* and *Language Modeling* attract many false positives (papers from other areas predicted as these), lowering precision, while simultaneously losing in-area papers to neighbouring labels (lower recall). We also see predictable adjacency confusions: a) *Generation* with *Summarization* with *Dialogue/QA*, b) *MT* with *Multilinguality*, c) *IR/text mining* with *Information Extraction*, d) *Semantics* with *Discourse/Pragmatics*, reflecting shared methods and datasets. Taken together, these patterns indicate that the current area set overaggregates meta roles and underspecifies boundaries for broad scopes; as a result, precision and recall degrade exactly where labels are vague or semantically similar in terms of topics covered. A hierarchical or primary plus secondary labeling (task/domain as primary; role/method such as “resources” “ethics” or “human-centered” as secondary) could be a way to reduce these systematic confusions and aid authors to select areas in a more systematic way.

7. Evaluation of Incorrectly Classified Data

Not only the analysis of the correctly classified papers offers insights, but also the incorrectly classified data points. In some instances, the models select the same area for one paper that does, however, not correspond to the one listed on Open Review.

For this, we filtered out the correctly classified papers and then compared the output of the models. As shown in Table 4, the *each category* prompting offers the most overlapping falsely classified papers, whereas *one-shot prompting* provides the least data points. This classification, therefore, is largely influenced by the performance of OpenAI-gpt-oss-20b.

Due to its low performance in the classification task, we filtered out the Hermes2Pro model for a second comparison, focusing on the two versions of the Qwen3-32B model and OpenAI-gpt-oss-20b.

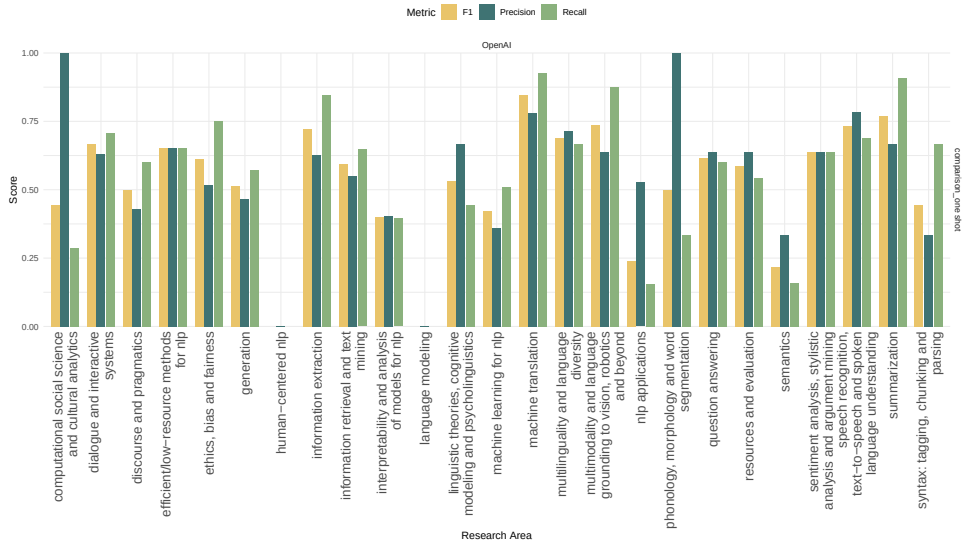

Figure 5: Performance of OpenAI-gpt-oss-20b in terms of precision, recall, and F1-score across the different areas at one-shot prompting.

The assigned label is the same for the three models in 114 cases for *each category*. This is equal to 20.43 % of the data set, the highest result of the five.

	All models	Qwen3–OpenAI
zero shot	51	96
one shot	50	91
three shot	59	101
five shot	63	104
each category	66	114

Table 4: Incorrectly classified but overlapping data

This comparison shows that the areas selected by the authors for the paper might not encompass the content correctly. This could be due to the difficulty of choosing an appropriate area. Table 4 illustrates an example of such a paper. The authors chose the broader category ”Resources and Evaluation”, whereas all LLMs selected the area ”Semantics”. OpenAI-gpt-oss-20b provides a reasoning for its selection, in which it is stated that ”the paper is about noun-noun compounds interpretation, semantic relations, linguistic study” and that ”the focus is semantics”. In this case, the area chosen by the LLMs might represent the content more accurately than the label chosen by the authors as might be inferred from the abstract (See Appendix B) of the paper which indicates a strong focus on semantics rather than providing a new resource.

It could also be interpreted as an indication that the areas offered by the ACL are not transparent, neither to the authors using the ARR area keywords to select an area or in general. The overlap of keywords across different areas might increase this effectas it becomes harder to distinguish the different areas if they encompass similar subtopics.

Article ID:	Rambelli_Chersoni...
Authors:	Resources and Evaluation
LLMs:	Semantics

Table 5: Example of a paper assigned a different area by all four LLMs than the one chosen by the authors. Metadata and abstract are in Appendix B.

8. Conclusion

In this paper, we compared the ability of different LLMs to accurately assign the thematic areas of ACL papers (ARR areas), using different prompting strategies (one-shot vs. few shot, prompting, inclusion of area-specific keywords). For this, we compiled a data set of 558 papers from ACL, EACL and NAACL, spanning from 2020 to 2025.

Our analyses of different prompting techniques do not fully conform to previous findings (Brown et al., 2020), which report higher accuracy with an increased number of in-context examples. In our experiments, the highest overall accuracy was achieved with one-shot prompting, suggesting that adding more examples does not necessarily benefit classification performance. Similarly, broadening the prompt to include all ARR areas in the dataset did not consistently improve model accuracy, with only one model showing performance gains. We also observed a substantial number of instances in which all models selected areas that differed from the authors' labels; in such cases, the

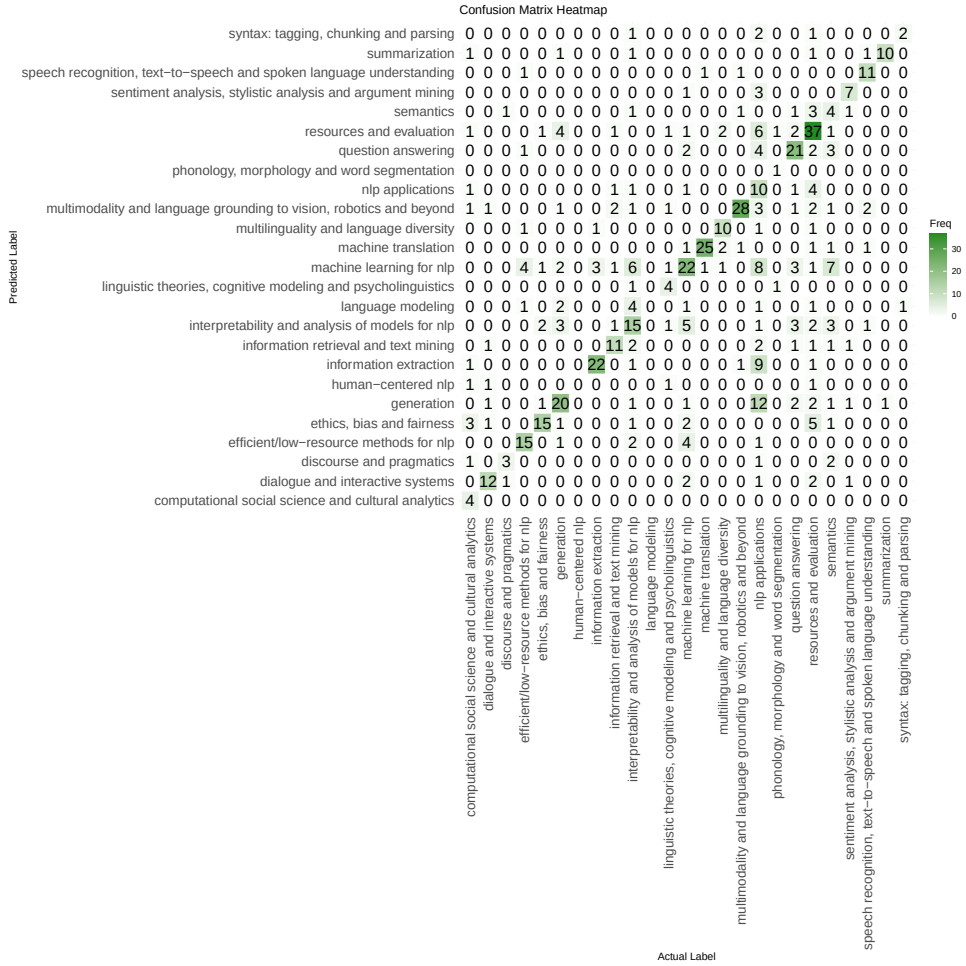

Figure 6: Confusion matrix for paper to area classification by the OpenAI-gpt-oss-20b model

models' predictions may be more thematically accurate than the authors' self-assigned areas. Overall, our results suggest that authors may experience difficulties in selecting thematically appropriate areas, potentially due to ambiguity in the area names and associated keywords.

Including the ARR's human-generated keywords, did not aid the models in the classification task. This might be due to the overlap of keywords in different areas. The keywords might not reflect the content of the papers in the area, contrasting with the semantic knowledge of the LLM.

The area labelling and keyword assignment is a dynamic process within ACL. Given our results, we suggest not only constant area revision aided by similar analyses to reflect the changing ACL landscape (as done in Rogers et al. (2023)), but also to install selection procedures of areas that reduce systematic confusions and aid authors to select areas in more systematic ways. An example could be hierarchically structured areas or the selection of a primary plus secondary labelling (e.g., task/domain as primary; role/method such as "resources" "ethics" or "human-centered" as secondary or vice versa).

9. Limitations

This work has several limitations. First, the number of LLMs evaluated was small. Although we tested various prompting strategies, including zero- and few-shot approaches with or without keywords, results may differ with other state-of-the-art or proprietary models. More advanced prompting and stronger LLMs could improve performance.

Second, the evaluation dataset was limited to 558 papers with uneven coverage across ARR research areas. This may have influenced performance differences and limits generalisability. Expanding the dataset and revising the ARR taxonomy, for example, using hierarchical or multi-label schemes, could reduce sampling bias, clarify label ambiguities, and better capture research complexities.

10. Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

11. Bibliographical References

2025a. ACL Rolling Review. https://aclrollingreview.org/. Accessed Sep 23, 2025.

2025b. ARR Area Keywords. https://aclrollingreview.org/areas. Accessed Sep 23, 2025.

2025. New Tracks at EMNLP 2025 and Their Relationship to ARR Tracks. https://2025.emnlp.org/track-changes/. Accessed Sep 23, 2025.

Raia Abu Ahmad, Ekaterina Borisova, and Georg Rehm. 2024a. Forc4cl: A fine-grained field of research classification and annotated dataset of nlp articles. In *International Conference on Language Resources and Evaluation*.

Raia Abu Ahmad, Ekaterina Borisova, and Georg Rehm. 2024b. Forc@nslp2024: Overview and insights from the field of research classification shared task. In *NSLP*.

Tom Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Fabio Dennstädt, Johannes Zink, Paul Martin Putora, Janna Hastings, and Nikola Cihoric. 2024. Title and Abstract Screening for Literature Reviews Using Large Language Models: An Exploratory Study in the Biomedical Domain. *Systematic Reviews*, 13(158).

Jennifer D'souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig. 2025. The Germeval 2025 2nd LLMs4Subjects Shared Task Dataset. Online dataset. Data managers and curators: Jennifer D'Souza (Data manager), Sameer Sadruddin (Data curator).

Clara Wan Ching Ho. 2025. UBFFM at the GermEval-2025 LLMs4Subjects Task: What if we take "You are an expert in subject indexing" seriously? In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 471–478, Hannover, Germany. HsH Applied Academics.

Lucas Joos, Daniel A. Keim, and Maximilian T. Fischer. 2024. Cutting Through the Clutter: The Potential of LLMs for Efficient Filtration in Systematic Literature Reviews. arXiv:2407.10652.

Taja Kuzman and Nikola Ljubešić. 2025. Llm teacher-student framework for text classification with no manually annotated data: A case study in iptc news topic classification. *IEEE access*.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of EMNLP*.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card.

Qwen Team. 2025. Qwen3 Technical Report.

Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Program chairs' report on peer review at acl 2023. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages xl–lxxv, Toronto, Canada. Association for Computational Linguistics.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv preprint arXiv:2402.07927*.

Samuel Schulhoff, Yuchen Liu, et al. 2024. A Systematic Survey of Prompt Engineering Techniques. *arXiv preprint arXiv:2406.06608*.

Parisa Shirali, Zahra Sarlak, and Ebrahim Ansari. 2025. Last Minute at the GermEval-2025 LLMs4Subjects Task: Few-Shot Contrastive Learning for Multilingual Multi-Label Classification. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 465–470, Hannover, Germany. HsH Applied Academics.

Tanja Säily, Jukka Suomela, Florent Perek, Jimena Jiménez Real, and Turo Vartiainen. 2025. Using Large Language Models to Enrich Corpus Metadata: The Case of Novels in COHA. In *46th Annual Conference of the International Computer Archive of Modern and Medieval English (ICAME 46)*, Vilnius, Lithuania.

Teknium, interstellarninja, theemozilla, karan4d, and huemin_art. 2024. Hermes-2-Pro-Llama-3-8B.

Kun Zhu, Lizi Liao, Yuxuan Gu, Lei Huang, Xiaocheng Feng, and Bing Qin. 2025. Context-aware hierarchical taxonomy generation for scientific papers via llm-guided multi-aspect clustering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15627–15645.

A. Prompt template

Role:
Act as a librarian and organize a collection of articles from the Association for Computational Linguistics (ACL), published between 2020 and 2025.
Objective:
Your task is to read, analyze, and organize a corpus of abstracts for papers from the following events:

- Association for Computational Linguistics (ACL)
- European Chapter of the Association for Computational Linguistics (EACL)
- Nations of America Chapter of the Association for Computational Linguistics (NAACL)

In general, the Association for Computational Linguistics focuses on the study of computational linguistics and natural language processing. Your goal is to create a comprehensive and structured database that facilitates search, and retrieval of these articles by researchers, and scientists.

Input:
You will be provided with instances from a dataset of astracts from the Association of Computational Linguistics. The dataset will consist of the abstracts of the original articles, along with some of their corresponding metadata:

- bibtex-style id
- title
- author(s)
- publication year
- event (ACL, EACL, NAACL)
- paper type (long or short)

Tasks:
A. Read and analyze the provided article to understand its core contribution and research context.
B. Identify the area of the paper. Areas describe the main idea of the paper and are used in the reviewing process for reviewer assignment. You should choose an area that: - accurately represents the content of the provided abstract

The area should exclusively be from this list of official areas:

1. Computational Social Science and Cultural Analytics
2. Dialogue and Interactive Systems
3. Discourse and Pragmatics
4. Efficient/Low-Resource Methods for NLP
5. Ethics, Bias, and Fairness
6. Generation
7. Human-Centered NLP
8. Information Extraction
9. Information Retrieval and Text Mining
10. Interpretability and Analysis of Models for NLP
11. Language Modeling
12. Linguistic Theories, Cognitive Modeling, and Psycholinguistics
13. Machine Learning for NLP
14. Machine Translation
15. Multilinguality and Language Diversity
16. Multimodality and Language Grounding to Vision, Robotics and Beyond
17. NLP Applications
18. Phonology, Morphology, and Word Segmentation
19. Question Answering
20. Resources and Evaluation
21. Semantics
22. Sentiment Analysis, Stylistic Analysis, and Argument Mining
23. Speech Recognition, Text-to-Speech and Spoken Language Understanding
24. Summarization
25. Syntax: Tagging, Chunking and Parsing

Do not use external sources such as the internet for this task.

Example Input:
ID: Sicilia_Gates_Alikhani_EACL_2024
Title: HumBEL: A Human-in-the-Loop Approach for Evaluating Demographic Factors of Language Models in Human-Machine Conversations
Authors: Anthony Sicilia, Jennifer Gates, Malihe Alikhani
Year: 2024
Event: eacl
Type: long
Abstract: While demographic factors like age and gender change the way people talk, and in particular, the way people talk to machines, [...].

Example Output (in YAML format):
article_id: "Sicilia_Gates_Alikhani_EACL_2024"
acl_area: "Dialogue and Interactive Systems"

Ensure that you output is a valid YAML file that follows the format provided above. No additional text output besides the YAML is required. Respect this rule and you will be rewarded accordingly.

B. Example of incorrectly classified paper

Article ID: 'Rambelli_Chersoni_Collacciani_Bolognesi_ACL_2024'
Title: 'Can Large Language Models Interpret Noun-Noun Compounds? A Linguistically-Motivated Study on Lexicalized and Novel Compounds'
Event: 'acl'
Year: '2024'
Author: 'Giulia Rambelli, Emmanuele Chersoni, Claudia Collacciani, Marianna Bolognesi'
Abstract: 'Noun-noun compounds interpretation is the task where a model is given one of such constructions, and it is asked to provide a paraphrase, making the semantic relation between the nouns explicit, as in carrot cake is "a cake made of carrots." Such a task requires the ability to understand the implicit structured representation of the compound meaning. In this paper, we test to what extent the recent Large Language Models can interpret the semantic relation between the constituents of lexicalized English compounds and whether they can abstract from such semantic knowledge to predict the semantic relation between the constituents of similar but novel compounds by relying on analogical comparisons (e.g., carrot dessert). We test both Surprisal metrics and prompt-based methods to see whether i.) they can correctly predict the relation between constituents, and ii.) the semantic representation of the relation is robust to paraphrasing. Using a dataset of lexicalized and annotated noun-noun compounds, we find that LLMs can infer some semantic relations better than others (with a preference for compounds involving concrete concepts). When challenged to perform abstractions and transfer their interpretations to semantically similar but novel compounds, LLMs show serious limitations.'