Zero-Shot Hierarchical Tagging of Short Technical Posts with Large Language Models

Nihad Badalov

Abstract

Short technical posts are difficult to organize: they are informal, context-dense, and rarely accompanied by task-specific labels. We study whether a general-purpose large language model can assign multiple paths from a user-defined hierarchy without fine-tuning. We evaluate a schema-constrained LLM on 80 English-language X posts sampled across eight areas of computer science, using blinded labels from one annotator and a frozen 155-node taxonomy. A 20-post development split selects between flat and indented hierarchy prompts and tunes zero-shot TF–IDF and sentence-embedding baselines; after exclusions, 58 held-out posts remain. Averaged over three held-out runs, the LLM reaches 0.622 ancestor-closed hierarchical F1 (95% bootstrap interval 0.558–0.686), compared with 0.574 for TF–IDF and 0.569 for MiniLM. Its exact-match rate is 0.264. Paired intervals for both hierarchical F1 and exact match include zero against both baselines, so the data do not establish superiority. All 180 held-out LLM responses are valid on the first attempt, but mean pairwise prediction-set Jaccard is 0.755 and taxonomy-gap detection is weak. An explicitly post-hoc, model-exposed author audit attributes many disagreements to reference omissions or corrections, highlighting label uncertainty rather than establishing improved accuracy. The study is a reproducible case analysis, not evidence of broad model dominance.

1 Introduction

Personal knowledge collections increasingly mix papers, documentation, blog posts, and short social-media posts. Tags make such collections searchable, but manual tagging is slow and inconsistent, especially when a collection uses a hierarchy rather than a flat vocabulary. Short posts amplify the difficulty: a few specialist phrases may imply a precise topic even when the broader field is never named.

This setting is a zero-shot hierarchical multi-label classification problem. The label set is fixed in advance, labels have parent–child relations, more than one label may apply, and there are no labeled training examples for fitting a task-specific classifier. Prior work shows that hierarchy-aware representations can improve zero-shot classification (Chalkidis et al., 2020; Liu et al., 2021; Paletto et al., 2024), while recent studies directly use large language models (LLMs) as hierarchical classifiers (Tabatabaei et al., 2025; Xia et al., 2025; Zhang et al., 2025). Those results do not settle the practical case considered here: an idiosyncratic 155-node taxonomy, terse technical posts, one author-annotator, and a command-line system whose output must remain machine-valid.

We therefore ask:

RQ1: How accurately does a schema-constrained LLM assign existing hierarchical tags compared with zero-shot lexical and embedding baselines?

RQ2: Does an indented tree representation improve development-set performance over a flat list of full paths?

RQ3: How stable are the LLM predictions across repeated runs, and how often does the structured-output pipeline require repair or produce invalid paths?

RQ4: Can proposed new tags identify genuine coverage gaps in the frozen taxonomy?

The contribution is intentionally narrow: a reproducible case study and evaluation harness, not a claim that one small corpus establishes general LLM superiority. Relative to the original prototype report, the present work freezes the taxonomy and corpus, introduces blinded reference labels, adds two zero-shot baselines, evaluates repeated runs, and publishes a sanitized artifact with explicit data boundaries.

2 Related Work

Text representations and zero-shot classification. TF–IDF remains a strong lexical reference point for retrieval and classification (Salton and Buckley, 1988; Manning et al., 2008). Dense sentence embeddings provide a semantic alternative by mapping texts and label descriptions to a shared vector space (Reimers and Gurevych, 2019). Zero-shot classification has also been framed as textual entailment (Yin et al., 2019). These families motivate our two deterministic baselines: cosine similarity between a post and lexical or dense label prototypes.

Hierarchical multi-label classification. Label hierarchies can support rare and unseen labels, but they also complicate evaluation because a near-miss within the correct branch differs from an unrelated prediction. Chalkidis et al. (2020) evaluate large-scale multi-label methods under few- and zero-shot labels; Liu et al. (2021) explicitly reasons over label dependencies and removes logically inconsistent predictions. Label augmentation can improve zero-shot hierarchical classification by enriching sparse label names (Paletto et al., 2024).

LLMs as hierarchical classifiers. LLM-based methods remove the need to train a task-specific classifier but introduce prompt, cost, validity, and stability questions. Text2Topic targets user-generated content with zero-shot capabilities (Wang et al., 2023). Industrial-scale evidence is mixed enough to warrant careful baselines and scope limits (Tabatabaei et al., 2025). Recent methods enrich label prototypes (Zhang et al., 2025), ensemble prompting strategies (Xia et al., 2025), or refine the taxonomy itself (Golde et al., 2026). Our study differs by fixing a small, user-created taxonomy and evaluating the full operational path from prompt serialization to parseable predictions.

3 Task and Taxonomy

Let x_i be the captured text of post i, and let $\mathcal{T}$ be a rooted forest represented by full slash-separated paths. The classifier returns a set $\hat{Y}_i \subseteq \mathcal{T}$ with at most three paths. The human reference is $Y_i \subseteq \mathcal{T}$, also capped at three non-redundant paths. "Non-redundant" means that an annotator does not select both a node and its descendant for the same semantic aspect.

The frozen taxonomy contains 155 nodes. Most lie under `cs`; a second root, `ai`, contains `ai/llms`. Source labels are preserved for audit, while machine paths are normalized with Unicode NFKD, ASCII transliteration, lowercase conversion, and underscore replacement. The export validator rejects duplicate normalized paths, missing parents, malformed indentation, and indentation jumps. The SHA-256 checksums in the artifact bind the manuscript to taxonomy version 1.0.0.

This taxonomy is deliberately imperfect. Some sampled topics, such as LLM evaluation or WebGPU, may lack an adequate leaf. Predictions used for the primary task must nevertheless be existing paths. A separate output field may propose a new path; proposals never mutate the taxonomy during the experiment.

4 Corpus Construction and Annotation

4.1 Sampling

We collected an 80-post X corpus on 20 August 2026 using the platform's Top-search surface in an authenticated browser. Search rankings are mutable; the study therefore records the exact observed candidate order privately and releases the selected post IDs publicly. Queries required English posts, excluded replies and retweets, used a lower date bound of 1 January 2023, and generally required at least 5 or 10 likes. We screened 468 observed candidates and retained ten posts from each stratum in Table 1. No author contributes more than two posts across the full corpus.

Eligibility required enough technical substance in the visible post text to support tagging. We excluded recruitment, generic engagement bait, pure product promotion without a technical claim, near-duplicate listicles, and posts whose meaning depended entirely on inaccessible quoted content or media. These judgments create selection bias and are not presented as a representative sample of X.

Table 1: Prespecified corpus strata. Each contributes ten posts.

Stratum	Representative search concepts
Compilers and languages	SSA, type systems, LLVM, garbage collection
Systems and concurrency	operating systems, io_uring, allocators, lock-free code
Hardware and GPUs	computer architecture, cache coherence, SIMD, memory hierarchy
Algorithms and theory	sorting, public-key cryptography, proof assistants, dynamic programming
Data and networks	distributed systems, database internals, network protocols
Software engineering	testing, hermetic builds, debugging, developer tools
Graphics and games	WebGPU, shaders, ray tracing, game physics
AI and ML systems	LLM inference, transformers, model evaluation, accelerators

4.2 Blinded reference annotation

The paper's sole author is also the sole human annotator. A local annotation application presents records in a deterministic randomized order and omits the sampling stratum, development/test assignment, and all model output. For each post, the annotator chooses one required primary path and up to two distinct secondary paths. The application rejects duplicate and ancestor–descendant-redundant selections. The annotator may exclude a record only if it is unavailable, non-English, duplicated, or technically insufficient, and must provide a reason.

The annotator separately marks whether the frozen taxonomy has a material gap and, if so, proposes a lowercase slash-separated path. This field is not used to score the primary existing-label task. Because there is one annotator, the study cannot estimate inter-annotator agreement; this is a central validity limitation, not a missing statistic.

4.3 Development and held-out split

We assign 20 posts to development and 60 to held-out test. Within each stratum, selected IDs are sorted by SHA-256 of a frozen seed concatenated with the post ID. The first three IDs in each of the first four strata and the first two in each remaining stratum form the development set. The split is hidden during annotation. Prompt representation, output thresholds, and baseline hyperparameters are selected only on development data; the test set is evaluated once after the configuration is frozen.

5 Systems Compared

5.1 LLM classifier

The classifier uses `gpt-5.4-mini` (OpenAI, 2026) through Codex CLI `0.151.0-alpha.7.1`. Each invocation is ephemeral and ignores user configuration and repository rules. It runs with a read-only sandbox, no repository dependency, low reasoning effort, and a strict JSON output schema. The prompt provides the post text, the complete taxonomy, the three-label cap, the non-redundancy rule, and an explicit instruction to use only existing paths for primary predictions.

Two taxonomy serializations are compared on development data:

1. **Indented hierarchy**: labels are rendered as a two-space-indented tree.
2. **Flat paths**: the same nodes are rendered as one full path per line.

The winning representation maximizes ancestor-closed hierarchical F1, with lower invalid-path rate and then higher exact match as tie-breakers. After selection, the prompt, schema, CLI flags, and model identifier are frozen. We run the selected LLM condition three times on the held-out set. A failed process or malformed response triggers a logged repair attempt up to a prespecified maximum of three attempts; raw responses remain private.

5.2 Zero-shot baselines

Both baselines construct one textual prototype per taxonomy node by replacing underscores with spaces and joining the node's ancestor labels. Cosine similarity ranks prototypes for a post. Development data selects a global score threshold and a maximum of one to three returned labels; ancestor–descendant duplicates are removed by retaining the more specific path.

TF–IDF. Post text and label prototypes share a word-and-bigram TF–IDF space. This baseline tests whether lexical overlap alone explains the LLM's performance (Salton and Buckley, 1988).

Sentence embeddings. The dense baseline uses `sentence-transformers/all-MiniLM-L6-v2` through FastEmbed 0.8.0. The 384-dimensional model has a 256-token input limit and is recorded by exact repository identifier in the artifact (sentence-transformers, 2021; Reimers and Gurevych, 2019). It tests semantic matching without generative inference or task-specific training.

6 Evaluation Protocol

For each post, exact match is $\mathbb{1}[Y_i = \hat{Y}_i]$. Sample F1 computes set F1 per post and averages it; micro F1 aggregates label–post decisions. To respect hierarchy, we close both sets under ancestors, $A(Y)$ and $A(\hat{Y})$, before computing hierarchical precision, recall, and F1. We also report the absolute difference between the deepest predicted and deepest reference path per post, averaged over posts, as depth error.

Operational metrics are the fraction of predicted paths absent from the frozen taxonomy, failed or repaired responses per invocation, and end-to-end retry counts. Stability across the three LLM runs is the mean pairwise Jaccard similarity of prediction sets per post. We do not treat model-reported confidence as calibrated probability.

All headline metrics receive 95% percentile bootstrap confidence intervals from 10,000 post-level resamples (Efron and Tibshirani, 1994). Paired bootstrap differences compare the frozen LLM condition with each baseline on hierarchical F1 and exact match. Resampling is stratified only in sensitivity analysis; the primary interval resamples held-out posts directly.

For RQ4, a human-marked taxonomy gap is the reference event. We report whether the model flags the gap and whether its proposed path is accepted as equivalent to the human proposal after author review. If fewer than five held-out gap cases remain, the analysis is descriptive only. No proposed node is inserted before primary scoring.

Only after all primary scores were frozen, the author conducted two exploratory reviews while exposed to model output. First, each model proposal on a true-positive gap occurrence was judged equivalent, not equivalent, or unclear relative to the human proposal. Second, every unique pair of held-out post and existing-taxonomy path predicted by at least one LLM run but absent from the frozen reference was reviewed once. Run identity and support were hidden until each decision was saved. Decisions distinguished an omitted co-label, a correction that should replace an original label, an acceptable alternative, an incorrect path, and an unclear case. Neither review changes the frozen reference or any primary score; only aggregate counts are reported.

7 Results

The annotation pass excluded five records and marked seven taxonomy gaps. Three exclusions occur in development and two in test, leaving 17 and 58 records respectively for classification metrics. Among all 80 assigned records, 53 receive one reference path, 16 receive two, six receive three, and the five excluded records receive none. The following results come exclusively from the frozen evaluation; the original prototype examples are not reused as evidence.

Table 2: Development-set selection on 17 eligible posts. Higher is better except invalid rate.

Condition	Hier. F1	Exact	Invalid rate
LLM, flat	0.738	0.471	0.000
LLM, indented	0.653	0.353	0.118
TF–IDF	0.540	0.294	0.000
MiniLM	0.608	0.235	0.000

Table 3: Held-out results on 58 eligible posts. Brackets give 95% percentile bootstrap intervals; lower depth error is better.

Condition	Exact	Sample F1	Micro F1	Hier. F1	Depth err.
LLM (three-run mean)	0.264 [0.167, 0.374]	0.429 [0.329, 0.532]	0.400 [0.311, 0.496]	0.622 [0.558, 0.686]	0.483 [0.333, 0.649]
TF–IDF	0.241 [0.138, 0.362]	0.351 [0.239, 0.463]	0.340 [0.237, 0.444]	0.574 [0.496, 0.652]	0.569 [0.379, 0.776]
MiniLM	0.224 [0.121, 0.328]	0.342 [0.236, 0.451]	0.340 [0.239, 0.443]	0.569 [0.491, 0.643]	0.672 [0.483, 0.862]

7.1 Development selection

Table 2 shows the prespecified development comparison. Flat full paths produce higher hierarchical F1 and exact match than indentation and no invalid records. The indented condition remains invalid after the allowed retries for two of 17 eligible records because it emits the case-sensitive non-path `Ai/LLMs`. We therefore freeze the flat representation. Development selection chooses a threshold of zero and one returned label for both baselines.

7.2 Held-out accuracy and hierarchy-aware performance

Table 3 reports point estimates and 95% post-bootstrap intervals on the 58 eligible held-out posts. The LLM entry averages metrics across its three runs; the bootstrap repeats that averaging within each resample. Its mean hierarchical precision and recall are 0.611 and 0.633. TF–IDF obtains 0.626 precision and 0.531 recall; MiniLM obtains 0.612 and 0.531. Thus, the LLM's hierarchical-F1 point estimate is higher mainly because it recovers more ancestor-closed reference paths, not because it is more precise than the lexical baseline.

The paired LLM-minus-TF–IDF difference is 0.048 in hierarchical F1 (95% interval −0.021 to 0.122) and 0.023 in exact match (−0.080 to 0.126). Against MiniLM, the corresponding differences are 0.054 (−0.025 to 0.134) and 0.040 (−0.069 to 0.149). Every paired interval includes zero. Consequently, this sample supports neither a claim of LLM superiority nor a claim of equivalence.

7.3 Validity, retries, and stability

Each held-out run processes all 60 assigned posts, including the two later excluded from accuracy scoring. All 180 invocations return schema-valid, taxonomy-valid output on the first attempt: there are no process failures, parsing failures, repairs, invalid paths, or retries. Output validity therefore differs sharply from semantic repeatability. Mean pairwise Jaccard similarity over eligible posts is 0.755, while exact prediction-set agreement is 0.609. Pairwise Jaccard values are 0.730, 0.790, and 0.744; exact agreements are 0.569, 0.672, and 0.586. The structured interface is operationally reliable in these runs, but the selected labels remain materially stochastic.

7.4 Taxonomy-gap analysis

Five held-out records are marked as genuine taxonomy gaps, meeting the prespecified minimum for quantitative reporting but still yielding highly uncertain estimates. The three LLM runs flag 8, 8, and 10 gaps. Their respective precision/recall/F1 values are 0.125/0.200/0.154, 0.250/0.400/0.308, and 0.100/0.200/0.133. None of the four true-positive proposal occurrences exactly matches the author's proposed path. In the post-hoc semantic review, one proposal is judged equivalent and three not equivalent; the equivalent proposal occurs in

run 2. Regardless of this adjudication, the detection results show both many false positives and most human gaps being missed.

7.5 Exploratory audit of model-only paths

Across 44 of the 58 eligible held-out posts, the three LLM runs produce 74 unique existing-taxonomy paths absent from the frozen reference. The post-hoc audit classifies 28 as missed valid co-labels, 21 as reference corrections, 12 as acceptable alternatives, eight as incorrect, and five as unclear. Of the 32 candidates selected by all three runs, 12 are judged missed co-labels, 12 reference corrections, five acceptable alternatives, one incorrect, and two unclear. These counts expose substantial uncertainty in a single-annotator reference, but they are not corrected accuracy estimates: the same author created the original labels and reviewed disagreements after seeing model suggestions.

8 Discussion

For RQ1, the LLM has the highest observed exact, sample-F1, micro-F1, and hierarchical-F1 point estimates, as well as the lowest depth-error estimate. The paired intervals, however, are compatible with a small disadvantage as well as a moderate advantage. On this corpus, schema-constrained prompting is promising but not demonstrated to outperform simple zero-shot retrieval baselines. The relatively large gap between flat micro F1 (0.400) and ancestor-closed F1 (0.622) indicates that many disagreements remain within a related hierarchy branch. Hierarchical credit makes those errors less severe, but does not make the leaf assignments correct.

The exploratory disagreement audit suggests that the frozen reference also omits or misstates many paths that the author later considers defensible. This finding cautions against treating exact agreement with one annotator as semantic ground truth. It does not justify revising the test set after model exposure or claiming higher model accuracy; an independent second annotation pass would be required for that inference.

For RQ2, flat paths clearly win the development comparison. Indentation supplies visible structure but also permits a surface-form error—`Ai/LLMs`—that strict path validation rejects. This is a development-set result rather than evidence that flat serialization is generally superior. It does show why prompt formats must be evaluated under the same machine-validity constraints as the deployed system.

For RQ3, structured output eliminates operational failures in the frozen held-out runs, yet repeated calls agree exactly on only about 61% of eligible posts. Parse validity and semantic stability are therefore distinct properties. A production personal-knowledge tool could exploit the reliable schema while routing low-consensus items to review, but confidence calibration and review policy are outside the present experiment.

For RQ4, the LLM is not an effective taxonomy-gap detector under this prompt. Even the best run identifies only two of five human gaps and produces six false alarms. Only one of four true-positive proposals is semantically equivalent to the author's missing concept. Semantic review can assess proposal wording after a gap is detected, but cannot repair the weak detection precision and recall. Prior work on label enrichment and taxonomy refinement suggests that dedicated descriptions or iterative refinement may help (Paletto et al., 2024; Golde et al., 2026); testing those interventions requires a new, independently evaluated protocol.

The main empirical lesson is therefore methodological. A valid JSON response is not necessarily a correct or stable classification, an ancestor-aware score should accompany rather than replace leaf-level metrics, and modest point-estimate gains on a small corpus should not be narrated as model superiority when paired uncertainty crosses zero.

9 Limitations and Ethics

The corpus is small, purposively stratified, restricted to English technical posts, and sampled from an opaque Top-search ranking. The human reference comes from one annotator who designed the taxonomy and system, introducing expertise benefits but also expectation bias. Blinding predictions and splits during the original annotation pass reduces one source of bias but does not establish annotation reliability. The post-hoc audits are weaker evidence still: the same author judged model disagreements after seeing the candidate paths, with no independent adjudicator or inter-annotator agreement measure. Their counts can diagnose possible

reference fragility but cannot be used to relabel the test set or recompute accuracy. The taxonomy is highly imbalanced: most nodes are under computer science, while the AI branch is shallow. Consequently, a model may be penalized for semantically reasonable distinctions that the hierarchy cannot express.

X content can be deleted, edited, or protected. The public artifact therefore contains post IDs and derived labels/predictions, not redistributed post text or profile data. Private captured text is used for annotation and reproducibility checks only. Rehydration remains subject to platform rules and post availability.

LLMs can reproduce biases from training data and can give plausible but unstable rationales (Bender et al., 2021). We score paths rather than treating model rationales as explanations. The experiment is also model- and date-specific: a pinned identifier improves reproducibility but does not guarantee that an externally hosted service remains bitwise stable.

10 Conclusion

In this 80-post case study, a schema-constrained LLM produces valid hierarchical tag sets reliably and attains a mean held-out hierarchical F1 of 0.622. Its point estimates exceed TF–IDF and MiniLM, but the paired bootstrap intervals do not establish an accuracy advantage. Flat taxonomy paths outperform indentation on development data. Across three repetitions, the mean pairwise Jaccard is 0.755 despite perfect first-attempt validity, and taxonomy-gap detection has low precision and recall. The post-hoc audit additionally reveals that a single frozen reference can itself be fragile. These findings support using LLM tagging as a reviewable component, not as an unquestioned replacement for deterministic baselines or human taxonomy maintenance. Larger corpora, multiple annotators, calibrated review thresholds, and evaluation on independently designed hierarchies are required before making broader claims.

Artifact Availability

The companion artifact distributed with the manuscript contains the frozen taxonomy, public corpus manifest, sampling protocol, prompts, schemas, evaluation code, sanitized gold labels and predictions, aggregate measurements, and exact dependency metadata. Code is licensed under MIT; derived labels and measurements are CC BY 4.0. Private post text, raw model responses, credentials, free-form notes, and annotation application state are excluded.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. doi: 10.1145/3442188.3445922.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of EMNLP*, pages 7503–7515, 2020. doi: 10.18653/v1/2020.emnlp-main.607. URL `https://aclanthology.org/2020.emnlp-main.607/`.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1994. doi: 10.1201/9780429246593.

Jonas Golde, Nicolaas Paul Jedema, RaviKiran Krishnan, and Phong Le. Hierarchical text classification with LLM-refined taxonomies. In *Proceedings of EACL*, pages 214–228, 2026. doi: 10.18653/v1/2026.eacl-long.10. URL `https://aclanthology.org/2026.eacl-long.10/`.

Hui Liu, Danqing Zhang, Bing Yin, and Xiaodan Zhu. Improving pretrained models for zero-shot multi-label text classification through reinforced label hierarchy reasoning. In *Proceedings of NAACL-HLT*, pages 1051–1062, 2021. doi: 10.18653/v1/2021.naacl-main.83. URL `https://aclanthology.org/2021.naacl-main.83/`.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Sch"utze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. URL https://nlp.stanford.edu/IR-book/.

OpenAI. GPT-5.4 mini model documentation. OpenAI developer documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.4-mini.

Lorenzo Paletto, Valerio Basile, and Roberto Esposito. Label augmentation for zero-shot hierarchical text classification. In *Proceedings of ACL*, pages 7697–7706, 2024. doi: 10.18653/v1/2024.acl-long.416. URL https://aclanthology.org/2024.acl-long.416/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988. doi: 10.1016/0306-4573(88)90021-0.

sentence-transformers. all-MiniLM-L6-v2 model card. Hugging Face model repository, commit 1110a243fdf4706b3f48f1d95db1a4f5529b4d41, 2021. URL https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

Seyed Amin Tabatabaei, Sarah Fancher, Michael Parsons, and Arian Askari. Can large language models serve as effective classifiers for hierarchical multi-label classification of scientific documents at industrial scale? In *Proceedings of COLING: Industry Track*, pages 163–174, 2025. URL https://aclanthology.org/2025.coling-industry.14/.

Fengjun Wang, Moran Beladev, Ofri Kleinfeld, Elina Frayerman, Tal Shachar, Eran Fainman, Karen Lastmann Assaraf, Sarai Mizrachi, and Benjamin Wang. Text2Topic: Multi-label text classification for efficient topic detection in user-generated content with zero-shot capabilities. In *Proceedings of the 2023 EMNLP Industry Track*, pages 93–103, 2023. doi: 10.18653/v1/2023.emnlp-industry.10. URL https://aclanthology.org/2023.emnlp-industry.10/.

Mingxuan Xia, Zhijie Jiang, Haobo Wang, Junbo Zhao, Tianlei Hu, and Gang Chen. Ensembling prompting strategies for zero-shot hierarchical text classification with large language models. In *Proceedings of EMNLP*, pages 18189–18208, 2025. doi: 10.18653/v1/2025.emnlp-main.918. URL https://aclanthology.org/2025.emnlp-main.918/.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of EMNLP-IJCNLP*, pages 3914–3923, 2019. doi: 10.18653/v1/D19-1404. URL https://aclanthology.org/D19-1404/.

Qian Zhang, Qinliang Su, Wei Zhu, and Pang Yachun. HierPrompt: Zero-shot hierarchical text classification with LLM-enhanced prototypes. In *Findings of EMNLP*, pages 3846–3859, 2025. doi: 10.18653/v1/2025.findings-emnlp.207. URL https://aclanthology.org/2025.findings-emnlp.207/.