AuthBench: A Large-Scale Multilingual Benchmark for Authorship Representation across Genres and Lengths

Anonymous ACL submission

Abstract

Authorship signals matter in settings where writing style carries identity: digital forensics, plagiarism analysis, account linking, misinformation investigation, and machine-generated text detection. Yet current authorship benchmarks remain fragmented, usually covering only a narrow language set, a single genre, or a limited document-length regime, which makes it difficult to assess whether modern representations truly generalize. We introduce AuthBench, a large-scale multilingual benchmark for authorship representation that is designed to make this evaluation broad, standardized, and realistic. AuthBench contains 428,150 documents written by 153,825 individuals across ten widely used languages, 9 primary genres, 66 fine-grained genres, and four document-length buckets. It supports two complementary tasks: *authorship attribution*, formulated as same-author retrieval and *authorship verification*, formulated as same-author binary decision. We benchmark 47 neural models and three non-neural baselines under a unified zero-shot protocol. Results show that authorship representation remains far from solved: the best retrieval model reaches only 0.258 Success@5, while the best verification model achieves 0.076 EER and 0.968 ROC-AUC. The leaderboard also reveals a meaningful task split, with different model families leading retrieval and verification, and large performance differences across languages, genres, and lengths. These findings position AuthBench not only as a new benchmark, but as a diagnostic resource for studying when and why authorship representations succeed or fail. We release AuthBench together with its evaluation toolkit anonymously at `https://anonymous.4open.science/r/AuthBench-712F/README.md`.

1 Introduction

Authorship representation (Huang et al., 2025; Habib et al., 2025; Tyo et al., 2022) examines whether a model can encode author-specific linguistic and stylistic features into reusable text representations, which is an important capability of current state-of-the-art (SOTA) models. This ability has a wide range of applications, including cybersecurity and digital forensics (Abbasi and Chen, 2008; Narayanan et al., 2012; Kumar et al., 2017; Caliskan-Islam et al., 2015), and underpins additional tasks such as plagiarism detection (Potthast et al., 2014; Barrón-Cedeño et al., 2013) and machine-generated text detection (Przystalski et al., 2025; Aityan et al., 2025; Al-Shaibani and Ahmed, 2026; Martinek and Bartuzi-Trokielewicz, 2024).

Driven by significant progress in large language models (LLMs) (Yang et al., 2025; Zhang et al., 2025; Grattafiori et al., 2024), SOTA models have become increasingly generalizable across downstream tasks, languages, and varying input lengths. However, existing resources still broaden authorship evaluation along only one or two axes at a time. Table 1 summarizes representative resources: classic corpora such as blogs and email collections are often monolingual and domain-specific; PAN established standardized evaluation protocols, but as a family of yearly shared tasks, each edition usually focuses on a narrow task or discourse setting; and more recent resources typically expand coverage along a single dimension, such as multilinguality, cross-discourse evaluation, cross-genre journalism, or document-length control. Consequently, prior work still lacks a unified benchmark that simultaneously offers multilingual coverage, broad genre diversity, explicit length control, and standardized support for both large-scale attribution and verification.

Benchmark	Year	Authors	Docs	Scope	Coverage			Evaluation		
					Multi-lingual	Multi-genre	Length bins	Standard splits	Attribution	Verification
Blogs50 (Zhang et al., 2018)	2006	50	66K	English blogs	✗	✗	✗	✗	✓	✗
MUD (adapted corpus) (Khan et al., 2021; Rivera-Soto et al., 2021)	2021	1.07M	321.7M	English Reddit	✗	✗	✗	✗	✗	✗
MARC (Keung et al., 2020)	2020	1.15M	1.26M	Multilingual reviews	✓	✗	✗	✓	✗	✗
PAN 2021 AV (PAN / Webis Group, 2021)	2021	278.2K	73.0K pairs	English fanfiction	✗	✗	✗	✓	✗	✓
PAN 2022 AV (PAN 2022 Authorship Verification Organizers, 2022)	2022	112	22.7K pairs	English cross-discourse	✗	✓	✗	✓	✗	✓
PAN 2023 AV (PAN 2023 Authorship Verification Organizers, 2023)	2023	112	18.5K pairs	English written/spoken	✗	✓	✗	✓	✗	✓
CROSSNEWS (Ma et al., 2025)	2025	2.26K	1.15M	English news + social	✗	✓	✗	✓	✓	✓
IMDb62 (Seroussi et al., 2014)	2011	62	62K	English reviews	✗	✗	✗	✗	✓	✗
CMCC (Goldstein-Stewart et al., 2008)	2008	21	756	English email/essay/blog/chat	✗	✓	✗	✗	✓	✗
Guardian (Stamatatos, 2013)	2013	13	1.0K	English opinion/book reviews	✗	✓	✗	✗	✓	✗
Enron Email (adapted corpus) (Klimt and Yang, 2004)	2004	∼150	∼500K	English email	✗	✗	✗	✗	✗	✗
SMAuC (adapted corpus) (Bevendorff et al., 2023)	2023	5.66M	3.36M	Scientific writing	✗	✓	✓	✗	✗	✗
AIDBench (identification benchmark) (Wen et al., 2024)	2024	3.25K	51.5K	English papers/e-mail/blogs/reviews	✗	✓	✗	✓	✗	✗
AuthBench (ours)	2026	153.8K	428.2K	Social, news, literature, blogs, etc.	✓	✓	✓	✓	✓	✓

Table 1: Comparison of AuthBench with representative prior authorship resources.

To address these limitations, we introduce AuthBench, a large-scale multilingual benchmark for authorship representation that supports attribution and verification across genres and document lengths. AuthBench combines author-linked text gathered through our own public-web crawling pipelines with carefully refined portions of existing datasets and benchmark resources, all standardized into a shared schema (see Appendix E). After quality filtering, deduplication, language auditing, and balanced sampling, the current release contains 428,150 documents of diverse lengths (short, medium, long, and extra-long), authored by 153,825 individuals across ten widely used languages (en, zh, hi, es, fr, ar, ru, de, ja, ko), spanning nine primary genres with 66 fine-grained sub-genres. AuthBench supports two complementary evaluation tasks: (i) *authorship attribution*, formulated as an open-world same-author retrieval task evaluated with ranking-based metrics (Success@K, Recall@K, nDCG@K, and MRR); and (ii) *authorship verification*, formulated as binary decision over query–candidate pairs and evaluated with EER and ROC-AUC. We release standardized dataset splits and a unified evaluation toolkit to facilitate reproducible and comprehensive evaluation.

We conduct a comprehensive zero-shot evaluation of 47 neural models and three non-neural baselines on AuthBench, spanning embedding models, base LLMs, and instruction-tuned variants. The leaderboard reveals a clear task split: `multilingual-e5-large` achieves the best overall retrieval performance (0.258 S@5, 0.254 R@5, 0.217 nDCG@5, 0.220 MRR), while `llama3.1-8b-instruct` achieves the best overall verification performance (0.076 EER, 0.968 ROC-AUC). Moreover, model quality still varies substantially across languages, document lengths, and genres, highlighting persistent challenges for robust authorship modeling.

2 Related Work

Authorship representation. Authorship representation (Huang et al., 2025; Habib et al., 2025; Tyo et al., 2022) seeks to identify persistent author-specific signals in written documents. The field is commonly structured around two core evaluation tasks: (i) authorship attribution, which links a document to its author and is traditionally studied as closed-set classification (Stamatatos, 2009; Neal et al., 2017) but is increasingly instantiated in representation-learning settings as open-world same-author retrieval over a candidate pool (Rivera-Soto et al., 2021; Wang and Iyyer, 2023); and (ii) authorship verification, which determines whether two documents were written by the same author (Neal et al., 2017; PAN 2022 Authorship Verification Organizers, 2022).

In parallel, authorship representation learning develops document embeddings that encode stylistic signals and can be used to support both attribution and verification at scale. Recent work therefore commonly reports retrieval metrics such as Recall@K or MRR for retrieval-style attribution, while standardized verification benchmarks emphasize ROC-based metrics such as AUC and representation-oriented analyses also report EER as a compact summary of the false-accept / false-reject trade-off (PAN 2022 Authorship Verification Organizers, 2022; PAN 2023 Authorship Verification Organizers, 2023; Wang and Iyyer, 2023).

Authorship Benchmark Development. Early authorship resources were instrumental for the field, but they were typically small-scale, monolingual, and tied to a narrow domain (Stamatatos, 2009; Neal et al., 2017). Classic datasets such as blog and email corpora (Schler et al., 2006a; Klimt and Yang, 2004) enabled important methodological progress, yet they offered limited coverage in language diversity, genre breadth, and explicit length control, and many were not released with broadly adopted benchmark splits. A second challenge is fragmentation: several resources widely used in modern work, including Blogs50, MUD, Enron, and SMAuC, are adapted from broader source collections rather than introduced as unified authorship benchmarks. PAN later established influential shared-task protocols for authorship analysis (Potthast et al., 2016; Bevendorff et al., 2020), but each edition typically focuses on a single task and discourse setting. More recent resources broaden the space only partially: MARC adds multilingual reviews (Keung et al., 2020), PAN 2022 and PAN 2023 study controlled cross-discourse verification in English (PAN 2022 Authorship Verification Organizers, 2022; PAN 2023 Authorship Verification Organizers, 2023), CROSSNEWS targets cross-genre journalism (Ma et al., 2025), SMAuC emphasizes scientific writing with length control (Bevendorff et al., 2023), and AIDBench centers on identification-style evaluations (Wen et al., 2024). What is still missing is a single benchmark that jointly supports multilingual, multi-genre, length-aware, standardized evaluation for both retrieval-style attribution and verification. AuthBench is designed to fill that gap.

3 AuthBench

In this section, we first describe the AuthBench construction in Section 3.1. Then we present the statistics of the AuthBench in Section 3.2.

3.1 Benchmark Construction

Figure 2 summarizes the AuthBench construction workflow, and Figure 3 gives a quick view of the final record format. AuthBench is built from two complementary inputs: author-linked text gathered through our public-web crawling pipelines, and selected existing datasets or benchmark resources that add useful languages, genres, or length regimes (Appendix E). We then standardize everything through five stages: Build & Normalization, Quality Filtering, Redundancy Reduction, Language Audit, and Bucket Balanced Sampling. Stage 1 maps every source into a shared schema, chunks overly long documents, and caps per-author document counts so heterogeneous materials become comparable without letting prolific authors dominate the benchmark.

The later stages turn that broad collection into a reliable evaluation benchmark. Quality Filtering removes noisy, low-information, and script-mismatched text; Redundancy Reduction removes exact duplicates, near-text overlaps, and optional near-author overlaps to reduce leakage; and the Language Audit verifies or retags suspicious labels. Bucket Balanced Sampling then applies hierarchical targets over language, genre, and length bucket before writing stratified train/dev/test splits, keeping the final release diverse, balanced, and reproducible. In short, the pipeline converts both crawled and inherited resources into a unified benchmark optimized for consistent authorship evaluation. Complete implementation-aligned details are provided in Appendix A.

3.2 Statistics of AuthBench

After construction, AuthBench comprises 428,150 documents authored by 153,825 individuals across 10 languages: English (en), Spanish (es), Chinese (zh), French (fr), German (de), Arabic (ar), Russian (ru), Japanese (ja), Korean (ko), and Hindi (hi). AuthBench spans 9 primary genres, including `social_media`, `literature`, `news`, `blog`, `media_reviews`, `poetry`, `ecommerce_reviews`,

Category	#Docs	Share
Overall		
Total	428,150	100.0%
Languages	10	–
Authors	153,825	–
Avg. docs per author	2.78	–
Genres (fine-grained)	66	–
By language		
English (en)	97,974	22.9%
Spanish (es)	33,395	7.8%
Chinese (zh)	55,368	12.9%
French (fr)	31,225	7.3%
German (de)	39,813	9.3%
Arabic (ar)	42,091	9.8%
Russian (ru)	66,084	15.4%
Japanese (ja)	21,494	5.0%
Korean (ko)	33,881	7.9%
Hindi (hi)	6,825	1.6%
By length		
Short	24,476	5.7%
Medium	211,910	49.5%
Long	184,447	43.1%
Extra-long	7,317	1.7%

Table 2: Statistics of our AuthBench.

`qna`, and `research_paper`. The distribution is dominated by `social_media` (40.9%), `literature` (30.0%), and `news` (17.2%), with the remaining six primary genres accounting for the final 11.9% of documents. Documents in AuthBench are further categorized into four token-length buckets: short (1–10 tokens), medium (11–100), long (101–500), and extra-long (>500). An overview of AuthBench statistics is presented in Table 2. The current split materialization contains 198,345 query documents and 229,805 candidate documents, partitioned into train/dev/test as 156,335/21,008/21,002 queries and 186,184/21,813/21,808 candidates. Additional detailed statistics are provided in Appendix D.

Figure 1 complements Table 2 with a compact visual summary of the benchmark profile. The language-level counts show that AuthBench remains broadly distributed across all ten languages rather than collapsing into a single dominant language, while the language–genre pie-chart reveals clear source-driven heterogeneity in genre coverage across languages. The token-length plot further shows that most documents concentrate in the medium-to-long range, with shorter texts and extra-long documents appearing less frequently. Together, these views highlight that AuthBench is broad in coverage but intentionally preserves realistic distributional variation across languages, genres, and lengths.

3.3 Pipeline Summary

4 Experiment Setup

Evaluation details. We embed queries and candidates with each model's encoder (or the final hidden states of LLM backbones) using a unified pipeline: tokenize each input, extract the last hidden states, pool into a single vector (mean pooling by default), and L2-normalize the resulting embedding. Similarities are computed with cosine similarity between normalized embeddings. We use full-sized within-split candidate pool; verification metrics are computed over all same-author and non-matching query–candidate pairs in that pool. The complete experiment setup is defined in Appendix A.7.

Models. We evaluate 47 neural models together with three non-neural baselines: a character 3–5 gram TF-IDF cosine baseline (Salton and Buckley, 1988), a lightweight stylometric n-gram baseline inspired by Koppel and Schler (2004), and a PPM-style compression baseline following Teahan and Harper (2003). These systems complement the neural leaderboard with strong lexical and compression-based reference points. Full model details are provided in Table 7 of Appendix A.7.

Metrics. AuthBench supports both authorship attribution and authorship verification tasks. For authorship attribution, we report Success@5 (S@5), Recall@5 (R@5), nDCG@5, and MRR; these respectively capture shortlist utility, coverage of multiple same-author targets, ranking quality near the top of the list, and the rank of the first correct same-author hit. For authorship verification, we compute EER and ROC-AUC; EER summarizes the balanced operating point where false acceptance and false rejection are equal, while ROC-AUC measures threshold-independent separability between same-author and different-author pairs. The main leaderboard reports

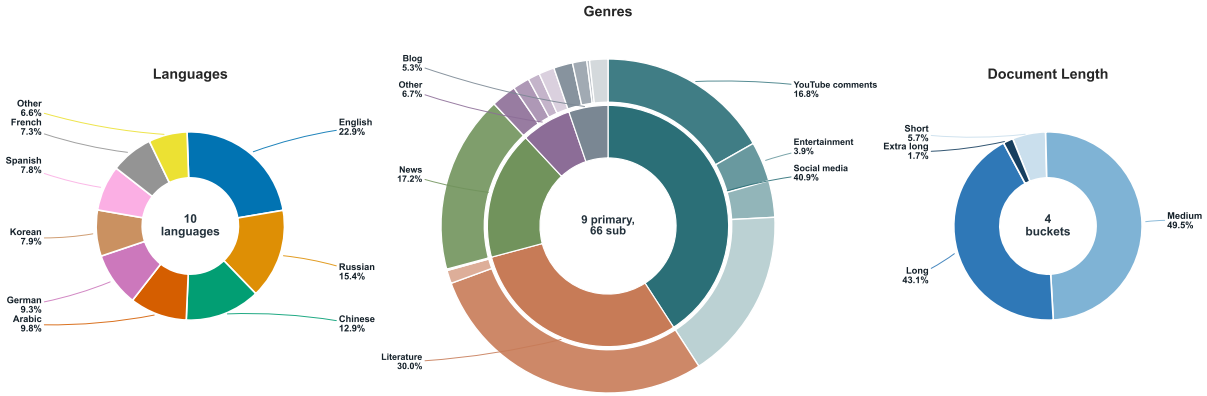

Figure 1: Overview of the current AuthBench profile for document proportion across languages, genres and document lengths.

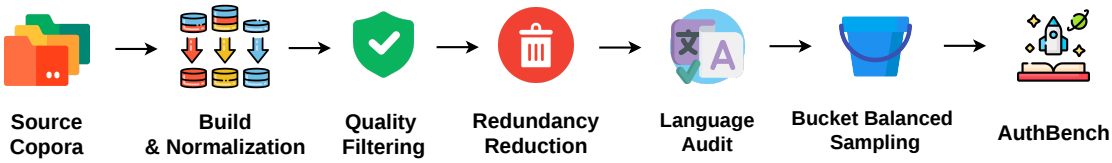

Figure 2: AuthBench construction pipeline. Full details are provided in Appendix A.

all six aggregate metrics, while the slice tables below focus on S@5 for compactness. Formal definitions of all evaluation metrics are provided in Appendix A.8.

5 Experimental Results

All results in this section are reported on the AuthBench test split, which contains 21,002 queries and 21,808 candidates. In this section, we first present the overall performance of SOTA models on AuthBench in Section 5.1. We then provide a detailed analysis of performance broken down by language, genre, and document length. Finally, in Section 5.5, we clarify the scope of the current zero-shot study and outline post-training as future work rather than as a reported main result.

5.1 Main Results

Finding 1: retrieval and verification favor different model families.

Table 3 shows a clear split between ranking and verification ability. For retrieval, the strongest model is `multilingual-e5-large`, reaching 0.258 S@5, 0.254 R@5, 0.217 nDCG@5, and 0.220 MRR. The top base LLMs are extremely close: `llama3-8b` and `llama3.1-8b` reach 0.254 and 0.253 S@5, respectively. This narrow gap suggests that large generative backbones already encode strong authorship cues, but specialized multilingual embedding models remain slightly better at turning those cues into stable nearest-neighbor structure.

Verification tells a different story. `llama3.1-8b-instruct` is the strongest verifier with 0.076 EER and 0.968 ROC-AUC, followed closely by `llama3-8b-instruct` and `llama3-8b`. In contrast, the best retrieval model is not the best verifier. This is an important takeaway for future work: authorship retrieval and authorship verification should not be treated as interchangeable probes of the same representation quality. Retrieval rewards global neighborhood structure, whereas verification also depends on pairwise calibration and decision-boundary sharpness.

The non-neural baselines remain informative but clearly behind the best neural systems. `tfidf` is the strongest lexical retrieval baseline at 0.197 S@5, while `ngram` is the strongest lexical verifier at 0.213 EER and 0.874 ROC-AUC. The gap to the neural frontier is still substantial, indicating that modern encoders capture authorial regularities beyond surface lexical overlap. At the same time, even the best neural scores remain far from ceiling, which suggests that robust authorship representation

```
"candidate_id": "doc_116326",
"author_id": "43b5e789...c39320",
"lang": "en",
"genre": "social_media/people",
"content": "How recently are these
things you found? People have pasts
... How would you like it?",
"source": "exorde",
"token_length": 156
```

```
"candidate_id": "doc_274629",
"author_id": "c423340f...30add",
"lang": "zh",
"genre": "news",
"content": "...",
"source": "babel_briefings",
"token_length": 74
```


Figure 3: Examples in our AuthBench.

is still an open problem rather than a solved byproduct of general text embedding quality.

5.2 Results by Language

Finding 2: performance varies sharply across languages, and the best model differs by language. Table 4 shows that there is no single universally best model across languages. `multilingual-e5-large` leads on Spanish, French, Japanese, Korean, and Russian; `qwen3-embedding-8b` is strongest on Chinese; `llama3-8b` is strongest on English; `llama3.1-8b` leads on German and Hindi; and the lexical `tfidf` baseline is narrowly strongest on Arabic. This pattern suggests that multilingual authorship representation is not just a matter of scaling one architecture, but of matching representation bias to the linguistic and genre composition of each language slice.

The spread in difficulty is also large. Chinese is comparatively easy, with `qwen3-embedding-8b` reaching 0.425 S@5, whereas Russian and Korean remain difficult even for the best systems, topping out at 0.190 and 0.196. These differences imply that benchmark-wide averages can hide substantial cross-lingual brittleness. Future research should therefore treat multilingual authorship modeling as a robustness problem, not simply as macro-averaged multilingual transfer.

A second notable trend is that multilingual embedding models dominate much of the non-English landscape. Their advantage over LLMs is especially visible in Chinese, Spanish, French, Japanese, Korean, and Russian. This suggests that explicit multilingual representation alignment remains highly valuable for authorship retrieval, and that future progress may come from improving language-sensitive embedding geometry rather than relying on larger generative models alone.

5.3 Results by Genre

Finding 3: genre difficulty varies dramatically, and different genres favor different architectures. Table 5 reveals large genre-dependent differences in difficulty. The easiest slice is `research_paper`, where `qwen3-embedding-8b` reaches 0.838 S@5, followed by `qna` at 0.682 and `poetry` at 0.667. At the other extreme, `media_reviews` and `ecommerce_reviews` are much harder, with best S@5 values of only 0.088 and 0.119. This gap indicates that authorship signals are much easier to recover in domains with stronger personal regularity or more stable discourse conventions than in short, noisy, or highly template-driven review settings.

The identity of the best model also changes sharply by genre. `gte-qwen2-7b-instruct` is strongest on `literature`; `qwen3-embedding-8b` is strongest on `media_reviews` and `research_paper`; `multilingual-e5-large` leads on `qna` and `social_media`; and the `llama3` family leads `blog`, `news`, `poetry`, and `ecommerce_reviews`. Rather than pointing to one dominant architecture, the table suggests that different model families are capturing different kinds of authorial evidence.

A useful way to interpret this pattern is through discourse structure. Embedding models appear especially strong in structured, information-dense settings such as `research_paper` and `qna`, where topical organization and repeated compositional habits may be easier to preserve in fixed-vector spaces. LLMs remain very competitive in more stylisti-

Model	Model Size	S@5 ↑	R@5 ↑	nDCG@5 ↑	MRR ↑	ROC-AUC ↑	EER ↓
LLMs (instruction-tuned)							
`llama3-8b-instruct`	8B	**0.251**	**0.247**	**0.206**	**0.209**	0.967	0.080
`llama3.1-8b-instruct`	8B	0.247	0.243	0.204	0.208	**0.968**	**0.076**
`qwen2.5-7b-instruct`	7.6B	0.207	0.203	0.168	0.172	0.964	0.083
`qwen3-4b-instruct`	4B	0.197	0.193	0.159	0.163	0.958	0.094
LLMs (base)							
`llama3-8b`	8B	**0.254**	**0.250**	**0.210**	**0.213**	**0.967**	**0.079**
`llama3.1-8b`	8B	0.253	0.249	0.209	0.212	0.967	0.080
`deepseek-llm-7b-base`	7B	0.220	0.216	0.180	0.184	0.954	0.105
`qwen3-4b`	4B	0.210	0.206	0.171	0.175	0.960	0.087
Embedding models (instruction-tuned)							
`e5-mistral-7b-instruct`	7.1B	**0.242**	**0.238**	**0.202**	**0.205**	0.955	0.096
`gte-qwen2-7b-instruct`	7.6B	0.240	0.235	0.198	0.202	**0.961**	**0.078**
Embedding models							
`multilingual-e5-large`	559.9M	**0.258**	**0.254**	**0.217**	**0.220**	0.920	0.157
`multilingual-e5-base`	278M	0.250	0.245	0.209	0.212	0.918	0.161
`qwen3-embedding-8b`	7.6B	0.241	0.237	0.201	0.203	0.940	0.135
`sfr-embedding-mistral`	7.1B	0.240	0.236	0.201	0.205	**0.956**	**0.096**
Lexical / non-neural baselines							
`tfidf`	–	**0.197**	**0.194**	**0.167**	**0.183**	0.837	0.219
`ngram`	–	0.170	0.167	0.145	0.149	**0.874**	**0.213**
`ppm`	–	0.157	0.156	0.137	0.142	0.792	0.298

Table 3: Main results on AuthBench (test split). The retrieval leaderboard now reports S@5, R@5, nDCG@5, and MRR; the verification leaderboard reports both ROC-AUC and EER. **Bold** entries mark the best result within each model group for a metric, and underlined entries mark the second best. The full 50-model evaluation, including all three non-neural baselines, is deferred to Appendix B.

cally expressive genres such as `blog`, `news`, and `poetry`, where broader contextual modeling may better preserve subtle stylistic signatures. Future authorship benchmarks and methods should therefore pay closer attention to genre as a first-class modeling variable rather than a secondary reporting slice.

5.4 Performance by Length

Finding 4: longer documents are markedly easier than short ones, but the best model still depends on length regime. Table 6 presents the updated S@5 breakdown across four document lengths. The strongest short-document model is `multilingual-e5-base` at 0.187 S@5, the strongest medium-document model is `multilingual-e5-large` at 0.220, and the strongest long and extra-long models are `llama3.1-8b` at 0.439 and 0.543, respectively.

The length trend is strong and intuitive: authorship retrieval becomes much easier as more text is available. Extra-long documents are the easiest regime overall, and the gain is especially pronounced for the LLMs. For example, `llama3.1-8b` rises from 0.135 S@5 on short documents to 0.543 on extra-long ones. This suggests that much of the remaining challenge in authorship modeling comes from sparse-evidence settings, where models must identify stable stylistic cues from very limited text.

The strongest model also changes with length. Short and medium documents favor embedding models, with `multilingual-e5-base` and `multilingual-e5-large` leading those buckets, whereas long and extra-long documents favor `llama3.1-8b`. This suggests a useful modeling hypothesis for future work: compact embedding models may be better at extracting robust local stylistic cues when evidence is scarce, while larger LLMs benefit more from longer contexts that expose higher-order discourse and syntactic habits.

Verification follows the same overall pattern. The best EER improves from 0.104 on `short` to 0.078 on `medium`, 0.060 on `long`, and 0.048 on `extra_long`. Short documents therefore remain the clearest bottleneck for both retrieval and verification. Advancing this regime will likely require methods explicitly designed for

Model	Model Size	ar	de	en	es	fr	hi	ja	ko	ru	zh
LLMs (instruction-tuned)											
`llama3-8b-instruct`	8B	**0.226**	0.244	**0.261**	**0.265**	**0.269**	0.289	**0.243**	0.173	0.164	**0.382**
`llama3.1-8b-instruct`	8B	0.211	**0.245**	0.259	0.265	0.265	**0.297**	0.229	**0.177**	**0.165**	0.373
`qwen2.5-7b-instruct`	7.6B	0.182	0.210	0.201	0.208	0.214	0.215	0.215	0.143	0.143	0.343
`qwen3-4b-instruct`	4B	0.187	0.195	0.187	0.203	0.198	0.229	0.212	0.143	0.143	0.304
LLMs (base)											
`deepseek-llm-7b-base`	7B	0.183	0.216	0.238	0.223	0.220	0.238	0.200	0.144	0.131	0.367
`llama3-8b`	8B	0.221	0.247	**0.270**	**0.272**	**0.270**	0.289	0.242	0.183	0.165	**0.380**
`llama3.1-8b`	8B	**0.225**	**0.249**	0.270	0.272	0.266	**0.297**	**0.246**	**0.186**	**0.173**	0.361
`qwen3-4b`	4B	0.189	0.217	0.204	0.213	0.214	0.227	0.220	0.142	0.147	0.336
Embedding models (instruction-tuned)											
`e5-mistral-7b-instruct`	7.1B	**0.223**	**0.241**	**0.237**	**0.254**	**0.250**	**0.272**	0.223	**0.164**	0.164	**0.396**
`gte-qwen2-7b-instruct`	7.6B	0.220	0.241	0.232	0.235	0.239	0.252	**0.249**	0.156	**0.171**	0.396
Embedding models											
`multilingual-e5-base`	278M	0.251	0.210	0.226	0.267	0.262	0.238	0.253	0.185	0.188	0.413
`multilingual-e5-large`	559.9M	**0.252**	0.221	0.231	**0.284**	**0.280**	**0.263**	**0.268**	**0.196**	**0.190**	0.422
`qwen3-embedding-8b`	7.6B	0.239	**0.242**	0.209	0.239	0.241	0.244	0.251	0.169	0.170	**0.425**
`sfr-embedding-mistral`	7.1B	0.221	0.236	**0.236**	0.255	0.248	0.258	0.214	0.162	0.162	0.397
Lexical / non-neural baselines											
`ngram`	–	0.238	0.137	0.126	0.162	0.166	0.244	0.160	0.168	0.132	0.267
`ppm`	–	0.229	0.140	**0.149**	0.190	**0.172**	0.255	0.113	0.103	0.118	0.184
`tfidf`	–	**0.254**	**0.148**	0.146	**0.190**	0.155	**0.280**	**0.226**	**0.184**	**0.154**	**0.346**

Table 4: Results by language (S@5). The updated slice table keeps a compact representative subset; complete language-wise S@5 and EER tables are reported in Appendix B.

low-evidence authorship signals, rather than simply scaling existing encoders.

5.5 Post-Training Outlook

In this paper, AuthBench serves as a large-scale pre-training evaluation benchmark for a broad set of authorship representation models, including embedding models, base LLMs, and instruction-tuned variants. This zero-shot comparison provides a useful foundation for future researchers to choose strong base models under different languages, genres, lengths, and task settings before performing post-training. We are also conducting follow-up analyses on post-training behavior and alternative authorship methods, and we plan to release a second version of the paper with a broader set of post-training results.

6 Conclusion

We introduced **AuthBench**, a large-scale benchmark for evaluating authorship representations across languages, genres, and document lengths under a unified retrieval-and-verification framework. Our zero-shot results show that authorship representation is still far from solved, that retrieval and verification favor different model families, and that robustness across languages, genres, and especially short documents remains the central challenge. We hope AuthBench provides a strong foundation for future work on more reliable and better calibrated authorship modeling in realistic multilingual settings.

Limitations

A key limitation of AuthBench is that language, genre, and document length are not fully orthogonal axes in the current release. Although we intentionally collected diverse sources and applied a robust filtering, deduplication, and auditing pipeline to reduce avoidable bias, some residual covariance across these axes is difficult to eliminate in practice. As a result, certain slice differences may still reflect source-level correlations rather than authorship difficulty alone. In addition, despite our leakage-reduction pipeline, some residual leakage may remain at scale. We view these as important targets for future benchmark refinement.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in

Model	Model Size	blog	ecommerce_reviews	literature	media_reviews	news	poetry	qna	research_paper	social_media
LLMs (instruction-tuned)										
llama3-8b-instruct	8B	0.213	0.107	**0.521**	**0.070**	**0.302**	**0.639**	**0.650**	0.769	**0.206**
llama3.1-8b-instruct	8B	**0.220**	**0.112**	0.518	0.035	0.293	0.583	0.618	**0.787**	0.204
qwen2.5-7b-instruct	7.6B	0.162	0.084	0.496	0.035	0.242	0.583	0.567	0.719	0.167
qwen3-4b-instruct	4B	0.149	0.080	0.462	0.053	0.227	0.611	0.548	0.688	0.160
LLMs (base)										
deepseek-llm-7b-base	7B	0.215	0.114	0.486	0.035	0.257	0.639	0.548	0.719	0.176
llama3-8b	8B	0.229	**0.119**	0.524	**0.070**	**0.311**	**0.667**	**0.631**	0.800	**0.207**
llama3.1-8b	8B	**0.232**	0.112	**0.532**	0.053	0.305	0.667	0.631	**0.812**	0.207
qwen3-4b	4B	0.166	0.084	0.485	0.035	0.241	0.639	0.561	0.738	0.171
Embedding models (instruction-tuned)										
e5-mistral-7b-instruct	7.1B	**0.202**	**0.110**	0.514	**0.053**	0.262	0.583	0.580	0.725	**0.207**
gte-qwen2-7b-instruct	7.6B	0.191	0.091	**0.538**	0.035	**0.281**	**0.611**	**0.586**	**0.781**	0.197
Embedding models										
multilingual-e5-base	278M	0.146	**0.107**	0.453	0.053	0.258	0.500	0.631	0.756	0.230
multilingual-e5-large	559.9M	0.160	0.100	0.479	0.053	**0.268**	0.444	**0.682**	0.769	**0.236**
qwen3-embedding-8b	7.6B	0.151	0.091	**0.536**	**0.088**	0.256	0.556	0.586	**0.838**	0.209
sfr-embedding-mistral	7.1B	**0.199**	0.105	0.512	0.053	0.257	**0.583**	0.567	0.756	0.207
Lexical / non-neural baselines										
ngram	–	0.066	0.050	0.289	0.018	0.161	0.278	0.325	0.431	0.168
ppm	–	**0.090**	**0.064**	0.230	0.000	0.149	0.250	0.420	**0.631**	0.153
tfidf	–	0.089	0.053	**0.323**	**0.035**	**0.169**	**0.444**	**0.459**	0.613	**0.198**

Table 5: Results by primary genre (S@5). Complete genre-wise S@5 and EER tables are reported in Appendix B.

cyberspace. *ACM Transactions on Information Systems*, 26(2):7:1–7:29.

Sergey K. Aityan, William Claster, Karthik Sai Emani, Sohni Rais, and Thy Tran. 2025. A lightweight approach to detection of AI-generated texts using stylometric features. *Preprint*, arXiv:2511.21744.

Maged S. Al-Shaibani and Moataz Ahmed. 2026. Arabic machine-generated text detection: Stylometric analysis and cross-model evaluation. *Expert Systems with Applications*, 305:130644.

arXiv. n.d. arxiv bulk data access. Website. Accessed: 2025-12-22.

barilan. n.d. Blog authorship corpus. Hugging Face Datasets. Accessed: 2025-12-22; canonical reference: (Schler et al., 2006b).

Alberto Barrón-Cedeño, Marta Vila, M. Antònia Martí, and Paolo Rosso. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4):917–947.

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. Shared tasks on authorship analysis at pan 2020. In *Advances in Information Retrieval, ECIR 2020*, volume 12036 of *Lecture Notes in Computer Science*, pages 508–516. Springer.

Janek Bevendorff, Philipp Sauer, Harrisen Scells, Benno Stein, and 1 others. 2023. SMAuC — the scientific multi-authorship corpus. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Aylin Caliskan-Islam, Richard Harang, Andrew Liu, Arvind Narayanan, Clare Voss, Fabian Yamaguchi, and Rachel Greenstadt. 2015. De-anonymizing programmers via code stylometry. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 255–270.

Cornell University. n.d. arxiv dataset. Kaggle Dataset. Accessed: 2025-12-22; see also (arXiv, n.d.).

S. Dhanwal and 1 others. 2020. An annotated dataset of discourse modes in hindi stories. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

Exorde Labs. 2024. Exorde social media (december 2024, week 1). Hugging Face Datasets. Accessed: 2025-12-22.

fengzhujoey. n.d. Douban dataset: Rating, reviews, side information. Kaggle Dataset. Accessed: 2025-12-22.

Jade Goldstein-Stewart, Kerri Goodwin, Roberta Sabin, and Ransom Winder. 2008. Creating and using a correlated corpus to glean communicative commonalities. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Google Developers. Youtube data api v3 documentation. https://developers.google.com/youtube/v3/docs. Accessed: 2026-03-15.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Model	Model Size	short	medium	long	extra_long
LLMs (instruction-tuned)					
llama3-8b-instruct	8B	**0.150**	**0.199**	**0.432**	**0.539**
llama3.1-8b-instruct	8B	0.142	0.196	0.427	0.531
qwen2.5-7b-instruct	7.6B	0.103	0.159	0.379	0.496
qwen3-4b-instruct	4B	0.081	0.150	0.368	0.488
LLMs (base)					
deepseek-llm-7b-base	7B	0.121	0.169	0.393	0.528
llama3-8b	8B	**0.150**	**0.202**	0.436	0.524
llama3.1-8b	8B	0.135	0.202	**0.439**	**0.543**
qwen3-4b	4B	0.101	0.161	0.385	0.496
Embedding models (instruction-tuned)					
e5-mistral-7b-instruct	7.1B	**0.161**	**0.196**	0.398	**0.524**
gte-qwen2-7b-instruct	7.6B	0.137	0.192	**0.409**	0.500
Embedding models					
multilingual-e5-base	278M	**0.187**	0.213	0.376	0.417
multilingual-e5-large	559.9M	0.182	**0.220**	0.392	0.476
qwen3-embedding-8b	7.6B	0.157	0.197	**0.395**	0.504
sfr-embedding-mistral	7.1B	0.168	0.193	0.395	**0.520**
Lexical / non-neural baselines					
ngram	–	0.151	0.136	0.271	0.331
ppm	–	0.079	0.134	0.255	0.240
tfidf	–	**0.157**	**0.164**	**0.300**	**0.437**

Table 6: Results by length bucket (S@5). Complete length-wise S@5 and EER tables are reported in Appendix B.

Nudrat Habib, Tosin Adewumi, Marcus Liwicki, and Elisa Barney. 2025. Trends and challenges in authorship analysis: A review of ml, dl, and llm approaches. *arXiv preprint arXiv:2505.15422*.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter*, 26(2):21–43.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Commonly released as MARC / Amazon Reviews Multi.

Aleem Khan, Elizabeth Fleming, Noah Schofield, Marcus Bishop, and Nicholas Andrews. 2021. A deep metric learning approach to account linking. *Preprint*, arXiv:2105.07263.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004*, volume 3201 of *Lecture Notes in Computer Science*, pages 217–226. Springer.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004)*, pages 489–495, Banff, Canada. ACM.

Srijan Kumar, Justin Cheng, Jure Leskovec, and V. S. Subrahmanian. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*.

Felix Leeb. 2023. Babel briefings. Hugging Face Datasets. Accessed: 2025-12-22.

Felix Leeb and Bernhard Schölkopf. 2024. A diverse multilingual news headlines dataset from around the world. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Dataset: Babel Briefings.

Marcus Ma, Duong Minh Le, Junmo Kang, Yao Dou, John Cadigan, Dayne Freitag, Alan Ritter, and Wei Xu. 2025. CROSSNEWS: A cross-genre authorship verification and attribution benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39.

Alicja Martinek and Ewelina Bartuzi-Trokielewicz. 2024. Detecting deepfakes and false ads through analysis of text and social engineering techniques. Manuscript. Text-based detection of AI-generated deepfake advertisement transcripts using linguistic and stylometric features.

mdanok. n.d. Arabic poetry dataset. Kaggle Dataset. Accessed: 2025-12-22.

mexwell. n.d. Amazon reviews multi. Kaggle Dataset. Accessed: 2025-12-22; cite (Keung et al., 2020) as the canonical dataset paper.

MIDAS Lab, IIIT-Delhi. n.d. Hindi discourse analysis dataset. GitHub Repository. Accessed: 2025-12-22; canonical reference: (Dhanwal et al., 2020).

Arvind Narayanan, Hristo S. Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. 2012. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy*, pages 300–314.

Tempestt J. Neal, Kalaivani Sundararajan, Ayesha Fatima, Yue Yan, Yang Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys*, 50(6).

PAN / Webis Group. 2021. Pan at clef 2021: Authorship verification. Official task website. Open-set cross-domain authorship verification task page.

PAN 2022 Authorship Verification Organizers. 2022. Overview of the authorship verification task at pan 2022. Task overview. Official overview used for the discourse-level statistics reported in the task.

PAN 2023 Authorship Verification Organizers. 2023. Overview of the authorship verification task at pan 2023. Task overview. Official overview covering the 2023 cross-discourse benchmark.

PleIAs. 2024a. French public domain books (french-pd-books). Hugging Face Datasets. Accessed: 2025-12-22.

PleIAs. 2024b. German public domain corpus (german-pd). Hugging Face Datasets. Accessed: 2025-12-22.

PleIAs. 2024c. Russian public domain corpus (russian-pd). Hugging Face Datasets. Accessed: 2025-12-22.

PleIAs. 2024d. Spanish public domain books (spanish-pd-books). Hugging Face Datasets. Accessed: 2025-12-22.

Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Winfried Lötzsch, Fabian Müller, Maike Elisa Müller, Robert Paßmann, Bernhard Reinke, Lucas Rettenmeier, Thomas Rometsch, Timo Sommer, Michael Träger, Sebastian Wilhelm, Shlomo Argamon, Moshe Koppel, Efstathios Stamatatos, and 2 others. 2016. Who wrote the web? revisiting influential author identification research applicable to information retrieval. In *Advances in Information Retrieval, ECIR 2016*, volume 9626 of *Lecture Notes in Computer Science*, pages 393–407. Springer.

Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, and Benno Stein. 2014. Overview of the 6th international competition on plagiarism detection. In *Working Notes Papers of the CLEF 2014 Evaluation Labs (CLEF 2014)*.

Project Gutenberg. Project gutenberg offline catalogs. `https://www.gutenberg.org/ebooks/offline_catalogs.html`. Accessed: 2026-03-15.

Karol Przystalski, Jan K. Argasiński, Iwona Grabska-Gradzińska, and Jeremi K. Ochab. 2025. Stylometry recognizes human and LLM-generated texts in short samples. *Preprint*, arXiv:2507.00838.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919. Association for Computational Linguistics.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006a. Effects of age and gender on blogging. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pages 199–205.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006b. Effects of age and gender on blogging. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*. Blog Authorship Corpus.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Stack Exchange Meta. 2025. Data dumps releases timeline, updates, and clarification. `https://meta.stackexchange.com/questions/396597/data-dumps-releases-timeline-updates-and-clarification`. Accessed: 2026-03-15.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Language and Politics*, 12(3):421–439. Includes the original Guardian corpus statistics used in later topic-bias studies.

William J. Teahan and David J. Harper. 2003. Using compression-based language models for text categorization. In W. Bruce Croft and John Lafferty, editors, *Language Modeling for Information Retrieval*, volume 13 of *The Information Retrieval Series*, pages 141–165. Springer, Dordrecht.

Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. 2022. On the state of the art in authorship attribution and authorship verification. *arXiv preprint arXiv:2209.06869*.

Anqi Wang and Mohit Iyyer. 2023. Can authorship representation learning capture stylistic similarity? *Transactions of the Association for Computational Linguistics*, 11:1431–1450.

Zichen Wen, Dadi Guo, and Huishuai Zhang. 2024. AIDBench: A benchmark for evaluating the authorship identification capability of large language models. *arXiv preprint*, arXiv:2411.13226.

Wikimedia Dumps. Wikisource dump index. `https://dumps.wikimedia.org/enwikisource/latest/`. Representative dump index for Wikisource; accessed: 2026-03-15.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

yuanchunhong. n.d. Xiaohongshu aigc comments (including posts). Kaggle Dataset. Accessed: 2025-12-22.

Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2742–2753, Brussels, Belgium. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A AuthBench Construction Details

This appendix provides the full construction specification for AuthBench, including Build & Normalization, Quality Filtering, Redundancy Reduction, Language Audit, and Bucket Balanced Sampling. Section 3.1 in the main paper contains a concise summary.

AuthBench is constructed through a five-stage pipeline: Build & Normalization, Quality Filtering, Redundancy Reduction, Language Audit, and Bucket Balanced Sampling. The pipeline is designed for large-scale ingestion while applying final benchmark size control only after Quality Filtering and Redundancy Reduction.

A.1 Stage 1: Build & Normalization - Source Ingestion and Normalization

Let $\mathcal{S} = \{S_1, \ldots, S_m\}$ denote the set of raw sources. Each source yields records with source-specific metadata, but the pipeline converts every surviving item into a unified document representation

$$x = (r,\ a,\ \ell,\ g,\ \sigma,\ t,\ L(t),\ k(t),\ \mathbf{m}),$$

where r is a source-level raw identifier, a is an anonymized author identifier, ℓ is language, g is the normalized genre, σ is source, t is document text, $L(t)$ is token length, $k(t)$ is the length bucket, and $\mathbf{m}$ contains optional metadata. Author identifiers are hashed deterministically as

$$a = \mathrm{SHA256}(\sigma : \texttt{raw_author}).$$

Token length is computed with a fixed tokenizer $T(\cdot)$ (here, `cl100k_base`):

$$L(t) = |T(t)|.$$

Length buckets are then assigned by

$$k(t) = \begin{cases} \texttt{short} & 1 \le L(t) \le 10, \\ \texttt{medium} & 11 \le L(t) \le 100, \\ \texttt{long} & 101 \le L(t) \le 500, \\ \texttt{extra_long} & L(t) > 500. \end{cases}$$

During ingestion, the builder may apply a bounded buffer shuffle to each dataset stream, which approximates random ordering without loading the full source into memory.

A.2 Stage 1: Build & Normalization - Chunking and Author Qualification

The first stage prepares a large author-qualified pool before any final size cap is enforced. If a raw document exceeds the chunking threshold, the pipeline segments it using paragraph and punctuation boundaries, with token-aware fallback splitting for very long sentences. Given a document d with text t, chunking produces

$$C(d) = \{t_1, \ldots, t_n\}, \qquad L(t_i) \le L_{\max},$$

where the pipeline uses configurable chunking parameters $(L_{\min}, L_{\text{target}}, L_{\max})$. Optional truncation is then applied by keeping the longest prefix that respects a specified token cap while preferring punctuation-aware segment boundaries.

After chunking, the builder applies a first-pass dirty-text filter (described below) and stores surviving documents in a bounded per-author reservoir. Let $\mathcal{D}_a$ be the documents observed for author a. Stage 1 keeps at most M documents per author in memory via reservoir replacement, with default $M = 5$. At finalization, authors with $|\mathcal{D}_a| < m$ are normally discarded, where the default target is $m = 3$ and a fallback minimum of 2 is allowed to recover sparse authors. Thus, the builder preferentially carries forward authors satisfying

$$m \le |\mathcal{D}_a| \le M,$$

with fallback admission for authors having exactly two clean documents. Stage 1 is intentionally permissive with respect to overall corpus size: it preserves as many clean, author-qualified items as possible for the later Quality Filtering and Bucket Balanced Sampling stages.

A.3 Stage 2: Quality Filtering

The second stage re-reads the author-qualified pool and performs stricter post-processing. First, the pipeline collapses letter-by-letter spacing artifacts such as "h e l l o" into normal words when the proportion or run length of single-letter alphabetic tokens is too high. Let $w = T_{\text{ws}}(t)$ denote whitespace-delimited tokens. The implementation measures

$$r_{\text{single}}(t) = \frac{1}{|w|} \sum_{j=1}^{|w|} \mathbf{1}[|w_j| = 1 \wedge \alpha(w_j)]\,,$$

$$m_{\text{single}}(t) = \max_j \max\{r : \forall k < r, \\ |w_{j+k}| = 1 \wedge \alpha(w_{j+k})\}.$$

If $r_{\text{single}}(t)$ or $m_{\text{single}}(t)$ exceeds a threshold, the pipeline attempts spacing collapse; if the cleaned text still exceeds the threshold, the document is dropped.

The post-filter then re-tokenizes the cleaned text and reruns rule-based dirty filtering. Let $w = T(t)$ be the token sequence. The dirty-text heuristics include:

- **Unique token ratio**

$$r_{\text{uniq}}(t) = \frac{|\text{unique}(w)|}{|w|},$$

  and the document is dropped if $r_{\text{uniq}}(t) < \tau_{\text{uniq}}$.

- **Symbol ratio**

$$r_{\text{sym}}(t) = \frac{\#\{\text{symbols in } t\}}{\max(|t|_{\text{chars}}, 1)},$$

  where public-domain sources use an additional consecutive-symbol check.

- **Maximum repeated-character run** $m_{\text{rep}}(t)$, the longest run of the same non-space character, which is dropped when $m_{\text{rep}}(t) > K_{\text{rep}}$.

- **Maximum consecutive-symbol run** $m_{\text{sym}}(t)$ for punctuation-heavy public-domain sources, dropped when $m_{\text{sym}}(t) > K_{\text{sym}}$.

The pipeline also applies an "untranslatable" filter that removes low-information or script-mismatched text. Define the alphabetic character ratio

$$r_\alpha(t) = \frac{\#\{\text{alphabetic characters in } t\}}{\#\{\text{non-space characters in } t\}},$$

$$r_{\text{tok}}(t) = \frac{1}{|w|} \sum_{j=1}^{|w|} \mathbf{1}[\alpha(w_j)],$$

where $\alpha(w_j)$ indicates that token w_j contains at least one alphabetic character. Let S_ℓ denote the expected Unicode script set for language ℓ. The script-match ratio is

$$r_{\text{script}}(t; \ell) = \frac{\#\{\text{letters in } t \text{ whose script} \in S_\ell\}}{\#\{\text{letters in } t\}}.$$

The post-filter rejects texts with low $r_\alpha(t)$, low $r_{\text{tok}}(t)$, excessive single-letter density, or low $r_{\text{script}}(t; \ell)$ once enough alphabetic characters are present. When script evidence is weak, the filter can also consult `langdetect`; if the detected language is incompatible with the target label, the document is removed at this stage.

A.4 Stage 3: Redundancy Reduction

After Quality Filtering, the pipeline removes redundancy in three passes.

Exact normalized-text duplicates. For each document, let $\widetilde{t} = N(t)$ denote case-folded text with canonicalized whitespace. We compute a stable exact-text key

$$h_{\text{exact}}(t) = \text{BLAKE2b}_{128}(N(t)),$$

and keep only the first document for each unique key.

Near-text duplicates. For documents with at least a minimum token count, we compute a 64-bit SimHash over unigram and bigram features. Let $\mathcal{F}(t)$ be the multiset of unigram and bigram features extracted from $N(t)$. The near-text signature is

$$h_{\text{near}}(t) = \text{SimHash}_{64}(\mathcal{F}(t)).$$

Candidate pairs are generated by LSH banding on the 64-bit signature. Two documents t and t' are considered near duplicates when

$$\text{Ham}\big(h_{\text{near}}(t), h_{\text{near}}(t')\big) \le \lfloor (1 - \theta_{\text{near}}) \cdot 64 \rfloor,$$

where θ_{near} is the configured similarity threshold. By default, near-text comparisons are restricted to the same language.

Near-author duplicates. The pipeline also supports optional author-profile redundancy checks to reduce residual cross-source aliasing. For author a, let $\mathcal{D}_a^{(p)}$ be the first p representative documents after sorting by document strength, where p is a small constant. The profile text is formed by concatenation

$$t_a^\star = \bigoplus_{d \in \mathcal{D}_a^{(p)}} N(t_d),$$

and the profile signature is $h_{\text{author}}(a) = \text{SimHash}_{64}(t_a^\star)$. Two authors a and a' are treated as near duplicates when

$$\text{Ham}\big(h_{\text{author}}(a), h_{\text{author}}(a')\big) \le \lfloor (1 - \theta_{\text{author}}) \cdot 64 \rfloor.$$

subject by default to same-language and cross-source constraints. When a conflict is found, the pipeline keeps the stronger author profile, prioritizing larger document count and then larger total token mass.

A.5 Stage 4: Language Audit

After Redundancy Reduction, the pipeline runs an automated language audit. For each document, it recomputes the script-match ratio $r_{\text{script}}(t;\ell)$ and marks a document as suspicious when the ratio is too low for its declared language. In addition, the pipeline samples up to a configurable number of documents for probabilistic language identification using `langdetect`. Let $\widehat{\ell}(t)$ be the top detected language and $p(\widehat{\ell}(t) \mid t)$ its confidence. A document is flagged as a high-confidence mismatch when

$$\widehat{\ell}(t) \neq \ell \quad \text{and} \quad p(\widehat{\ell}(t) \mid t) \geq \tau_{\text{conf}}.$$

By default, such documents are retagged rather than dropped; optional stricter settings can discard them. The audit records suspicious cases for targeted manual review together with summary counts such as mismatch rate, low-script rate, and retagged language pairs.

A.6 Stage 5: Bucket Balanced Sampling

Let $\mathcal{D}^\star$ denote the audited document pool. The final benchmark target Q is applied only at this stage. The pipeline first groups documents by language and assigns a target

$$Q_\ell = \text{round}(Q\, p_\ell),$$

where p_ℓ is the configured language prior after renormalization over the languages that are actually available. Within language ℓ, genre targets are computed as

$$Q_{\ell,g} = \text{round}(Q_\ell \rho_{\ell,g}),$$

and within each language–genre slice the length-bucket targets are

$$Q_{\ell,g,k} = \text{round}(Q_{\ell,g}\beta_k),$$

where $\rho_{\ell,g}$ is the configured genre weight for language ℓ and β_k is the global length-bucket prior. Sampling is performed without replacement. If a bucket (ℓ, g, k) is underfull, the pipeline first redistributes the deficit to leftover documents from the same language and genre, and then uses a spill pool formed from remaining language-matched documents. This preserves the language target as closely as possible while relaxing finer-grained quotas only when necessary.

Split construction. Stage 5 also performs the final document-level stratified split. Documents are first grouped by language and then bucketed by

$$b(d) = (g(d), k(d)).$$

For each language-specific bucket $B_{\ell,g,k}$, the pipeline applies deterministic shuffling with a fixed seed and allocates integer split counts according to the requested train/dev/test ratios. If r_s is the desired ratio for split $s \in \{\texttt{train}, \texttt{dev}, \texttt{test}\}$, the raw allocation is $|B_{\ell,g,k}|\, r_s$, floored to integers, and any leftover documents are assigned by largest fractional remainder. This preserves genre and length composition within each language while keeping the split procedure deterministic.

Because exact and near-text redundancy reduction are applied globally before splitting, duplicate leakage across train/dev/test is reduced by construction. The current benchmark does *not* enforce author-disjoint splits; instead, it prioritizes balanced document distributions for downstream retrieval and verification within each split. Within each split, authors with at least two documents contribute 1–2 candidate documents and 1–2 query documents, and each query is paired with all candidate documents from the same author in the benchmark ground-truth records.

Overall, the pipeline yields a diverse benchmark with explicit control over language, genre, and length, while combining stream-safe ingestion, rule-based cleaning, multi-level redundancy reduction, automated language auditing, and deterministic stratified finalization.

A.7 Evaluation Protocol

For retrieval, each test document is used as a query and ranked against a candidate pool drawn from the same split. For verification, we evaluate on labeled query–candidate pairs derived from the same within-split pools, computing both EER and ROC-AUC. By default,

every non-matching candidate in the pool is treated as a negative for a given query; the toolkit also supports deterministic negative sampling for faster ablations. All results are computed without task-specific fine-tuning to ensure comparability across model families. We report both aggregate performance and mandatory stratified breakdowns by language, genre, and length. Length buckets follow the benchmark definition: short (1–10 tokens), medium (11–100), long (101–500), and extra_long (¿500).

Embedding extraction for LLM-based models. For base and instruction-tuned LLMs that do not expose a dedicated sentence-embedding head, we derive a single document vector from the final hidden states. Given an input document $x = (x_1, \ldots, x_n)$, let

$$H^{(L)}(x) = \left[h_1^{(L)}, \ldots, h_n^{(L)}\right], \qquad h_i^{(L)} \in \mathbb{R}^d,$$

denote the last-layer token representations over the non-padding positions. Our default pooled representation is the masked mean

$$\bar{h}(x) = \frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} m_i \, h_i^{(L)},$$

where $m_i \in \{0, 1\}$ indicates whether token x_i is a valid (non-padding) token. We then L2-normalize the pooled vector,

$$e(x) = \frac{\bar{h}(x)}{\|\bar{h}(x)\|_2},$$

and use $e(x)$ as the document embedding for both retrieval and verification. Similarity between a query q and candidate c is then computed as cosine similarity, which reduces to a dot product after normalization:

$$s(q, c) = e(q)^\top e(c).$$

This is the default path used to obtain embeddings from generic LLM backbones in our unified pipeline. When a model family provides its own native embedding interface or explicitly recommended pooling rule (e.g., special-token or last-token pooling), we follow that model-specific implementation instead.

A.8 Metrics

This metric design follows standard authorship-analysis practice while making our retrieval formulation explicit: classical authorship attribution is often closed-set classification (Stamatatos, 2009; Neal et al., 2017), recent representation-learning work frequently evaluates open-world same-author retrieval with ranking metrics (Rivera-Soto et al., 2021; Wang and Iyyer, 2023), and authorship verification remains the pairwise same-author decision problem typically assessed with ROC-based measures (PAN 2022 Authorship Verification Organizers, 2022; PAN 2023 Authorship Verification Organizers, 2023). Let q be a query with candidate set C_q and positives $P_q \subset C_q$. Let π_q be the ranking of candidates by similarity. We define the binary relevance at rank i as

$$\mathrm{rel}_i = \begin{cases} 1 & \text{if } \pi_q(i) \in P_q, \\ 0 & \text{otherwise.} \end{cases}$$

For a cutoff K (we report $K = 5$ in the main paper), we compute:

$$\mathrm{Recall@}K(q) = \frac{1}{|P_q|} \sum_{i=1}^{K} \mathrm{rel}_i$$

$$\mathrm{DCG@}K(q) = \sum_{i=1}^{K} \frac{\mathrm{rel}_i}{\log_2(i+1)}$$

$$\mathrm{IDCG@}K(q) = \sum_{i=1}^{\min(K, |P_q|)} \frac{1}{\log_2(i+1)},$$

$$\mathrm{nDCG@}K(q) = \frac{\mathrm{DCG@}K(q)}{\mathrm{IDCG@}K(q)}$$

$$\mathrm{RR}(q) = \begin{cases} \frac{1}{\min\{i : \mathrm{rel}_i = 1\}} & \exists i \in C_q : \mathrm{rel}_i = 1, \\ 0 & \text{otherwise.} \end{cases}$$

We define Success@K as:

$$\mathrm{Success@}K(q) = \mathbf{1}\left[\sum_{i=1}^{K} \mathrm{rel}_i > 0\right].$$

We report S@5, R@5, nDCG@5, and the macro-average mean reciprocal rank (MRR) over queries.

For authorship verification, each query–candidate pair is assigned a similarity score $s \in \mathbb{R}$ and classified using a threshold τ. Let $\mathcal{P}$ and $\mathcal{N}$ denote the sets of positive and negative pairs, respectively. The false acceptance

rate (FAR) and false rejection rate (FRR) at threshold τ are defined as

$$\mathrm{FAR}(\tau) = \frac{1}{|\mathcal{N}|} \sum_{(q,c)\in\mathcal{N}} \mathbf{1}[s(q,c) \geq \tau], \quad (1)$$

$$\mathrm{FRR}(\tau) = \frac{1}{|\mathcal{P}|} \sum_{(q,c)\in\mathcal{P}} \mathbf{1}[s(q,c) < \tau]. \quad (2)$$

The Equal Error Rate (EER) is defined as the operating point where the two error rates are equal:

$$\mathrm{EER} = \mathrm{FAR}(\tau^*) = \mathrm{FRR}(\tau^*),$$

$$\text{where } \tau^* = \arg\min_{\tau} |\mathrm{FAR}(\tau) - \mathrm{FRR}(\tau)|.$$

The ROC curve traces $\mathrm{TPR}(\tau) = 1 - \mathrm{FRR}(\tau)$ against $\mathrm{FAR}(\tau)$ as τ varies. We summarize this curve with the area under the ROC curve (ROC-AUC):

$$s^+ = s(q, c^+), \qquad s^- = s(q', c^-),$$

$$\text{ROC-AUC} = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{\substack{(q,c^+)\in\mathcal{P} \\ (q',c^-)\in\mathcal{N}}} \left(\mathbf{1}[s^+ > s^-] + \tfrac{1}{2}\mathbf{1}[s^+ = s^-]\right).$$

Higher ROC-AUC is better.

Why these metrics? Authorship attribution in AuthBench is an open-world retrieval problem with potentially multiple relevant candidates per query. Success@5 captures whether a model can surface at least one correct same-author document in a short analyst-facing shortlist. Recall@5 measures how much of the relevant same-author set is recovered within the top of the ranking. nDCG@5 complements these hit-based measures by rewarding correct documents that appear earlier and by accounting for queries with more than one relevant candidate. MRR adds a first-hit view of ranking quality, which is useful when only the first correct same-author retrieval matters operationally.

Authorship verification is a pairwise same-author decision task. EER is useful when a single balanced operating point is desired because it directly summarizes the trade-off between false accepts and false rejects. ROC-AUC complements EER by measuring score separability independent of any one threshold, which is important when raw similarity scales differ across models. We therefore use EER for an interpretable operating-point summary and ROC-AUC for threshold-independent discrimination, consistent with recent authorship verification practice (PAN 2022 Authorship Verification Organizers, 2022; PAN 2023 Authorship Verification Organizers, 2023; Wang and Iyyer, 2023).

A.9 Models Evaluated

Non-neural baseline details. `tfidf` represents each document with scikit-learn character 3–5 gram TF-IDF features and ranks candidates by cosine similarity, following the standard vector-space term-weighting formulation of Salton and Buckley (1988). `ngram` is a lightweight stylometric baseline inspired by the feature-based authorship verification line of Koppel and Schler (2004): our implementation combines hashed character 3–5 grams, hashed word 1–2 grams, and a small set of surface cues (length, punctuation, digits, capitalization, whitespace, and mean token length), then fits a train-split linear pair calibrator to produce same-author scores. `ppm` follows the compression-based language-modeling view of Teahan and Harper (2003): we approximate a fixed-order character PPM scorer with hashed character counts, derive symmetric query–candidate cross-entropy features, and fit a train-split linear calibrator. For `ngram` and `ppm`, these are scalable benchmark implementations inspired by the cited methods rather than exact historical reimplementations.

B Full Results Tables

This appendix reports the full zero-shot result tables for all 47 neural models and the three non-neural baselines evaluated on AuthBench. Table 8 through Table 14 provide the complete benchmark breakdown across overall, language, genre, and length settings, while Figures 4–9 provide metric-wise visual summaries.

B.1 Overall Leaderboard Full Results

This subsection reports the complete zero-shot leaderboard across all evaluated model families on the AuthBench test split.

B.2 Overall Metric Bar Charts

These figures provide metric-wise visual summaries of the full leaderboard and make the relative spread between model families easier to compare at a glance.

Table 7: Models and baselines evaluated in AuthBench. We list every system appearing in the updated leaderboard together with its model size and Hugging Face repository or baseline implementation note.

Model	**Model Size**	**Repository / Implementation**
LLMs (instruction-tuned)		
llama3-8b-instruct	8B	meta-llama/Meta-Llama-3-8B-Instruct
llama3.1-8b-instruct	8B	meta-llama/Llama-3.1-8B-Instruct
qwen2.5-3b-instruct	3.1B	Qwen/Qwen2.5-3B-Instruct
qwen2.5-7b-instruct	7.6B	Qwen/Qwen2.5-7B-Instruct
deepseek-llm-7b-chat	7B	deepseek-ai/deepseek-llm-7b-chat
qwen3-4b-instruct	4B	Qwen/Qwen3-4B-Instruct-2507
LLMs (base)		
llama3-8b	8B	meta-llama/Meta-Llama-3-8B
llama3.1-8b	8B	meta-llama/Llama-3.1-8B
deepseek-llm-7b-base	7B	deepseek-ai/deepseek-llm-7b-base
qwen2.5-3b	3.1B	Qwen/Qwen2.5-3B
qwen3-4b	4B	Qwen/Qwen3-4B
Embedding models (instruction-tuned)		
e5-mistral-7b-instruct	7.1B	intfloat/e5-mistral-7b-instruct
gte-qwen2-7b-instruct	7.6B	Alibaba-NLP/gte-Qwen2-7B-instruct
Embedding models		
multilingual-e5-large	559.9M	intfloat/multilingual-e5-large
multilingual-e5-base	278M	intfloat/multilingual-e5-base
qwen3-embedding-8b	7.6B	Qwen/Qwen3-Embedding-8B
sfr-embedding-mistral	7.1B	Salesforce/SFR-Embedding-Mistral
qwen3-embedding-4b	4B	Qwen/Qwen3-Embedding-4B
snowflake-arctic-embed-l-v2	567.8M	Snowflake/snowflake-arctic-embed-l-v2.0
qwen3-embedding-0.6b	595.8M	Qwen/Qwen3-Embedding-0.6B
e5-large-v2	335.1M	intfloat/e5-large-v2
e5-base-v2	109M	intfloat/e5-base-v2
facebook-contriever	110M	facebook/contriever
gte-large-en-v1.5	409M	Alibaba-NLP/gte-large-en-v1.5
bge-m3	567M	BAAI/bge-m3
bge-large-en-v1.5	335M	BAAI/bge-large-en-v1.5
e5-small-v2	33M	intfloat/e5-small-v2
gte-large	335.1M	thenlper/gte-large
mxbai-embed-large-v1	335.1M	mixedbread-ai/mxbai-embed-large-v1
bge-base-en-v1.5	109.5M	BAAI/bge-base-en-v1.5
gte-base	109.5M	thenlper/gte-base
bge-base-zh-v1.5	102M	BAAI/bge-base-zh-v1.5
facebook-contriever-msmarco	110M	facebook/contriever-msmarco
bge-large-zh-v1.5	326M	BAAI/bge-large-zh-v1.5
distiluse-base-multilingual-cased-v2	134.7M	sentence-transformers/distiluse-base-multilingual-cased-v2
all-roberta-large-v1	355.4M	sentence-transformers/all-roberta-large-v1
all-mpnet-base-v2	109.5M	sentence-transformers/all-mpnet-base-v2
bge-small-en-v1.5	33.4M	BAAI/bge-small-en-v1.5
all-minilm-l12-v2	33.4M	sentence-transformers/all-MiniLM-L12-v2
paraphrase-mpnet-base-v2	109.5M	sentence-transformers/paraphrase-mpnet-base-v2
bert-base-uncased	110.1M	bert-base-uncased
all-minilm-l6-v2	22.7M	sentence-transformers/all-MiniLM-L6-v2
paraphrase-multilingual-mpnet-base-v2	278M	sentence-transformers/paraphrase-multilingual-mpnet-base-v2
msmarco-distilbert-base-v4	66.4M	sentence-transformers/msmarco-distilbert-base-v4
allenai-specter	110M	allenai/specter
jina-embeddings-v2-small-en	33M	jinaai/jina-embeddings-v2-small-en
jina-embeddings-v2-base-en	137.4M	jinaai/jina-embeddings-v2-base-en
Lexical / non-neural baselines		
tfidf	–	scikit-learn character 3–5 gram TF-IDF cosine baseline
ngram	–	hashed character/word n-gram stylometric baseline with train-split calibrator
ppm	–	fixed-order hashed character language-model approximation of PPM-style scoring

Table 8: Updated overall zero-shot results on AuthBench. Authorship attribution is evaluated with Success@5 (S@5), Recall@5 (R@5), nDCG@5, and MRR (higher is better). Authorship verification is evaluated with ROC-AUC (higher is better) and EER (lower is better). All 47 neural models and the three non-neural baselines are included.

Model	**Model Size**	**S@5 ↑**	**R@5 ↑**	**nDCG@5 ↑**	**MRR ↑**	**ROC-AUC ↑**	**EER ↓**
LLMs (instruction-tuned)							
`llama3-8b-instruct`	8B	**0.251**	**0.247**	**0.206**	**0.209**	0.967	0.080
`llama3.1-8b-instruct`	8B	0.247	0.243	0.204	0.208	**0.968**	**0.076**
`qwen2.5-3b-instruct`	3.1B	0.213	0.209	0.172	0.176	0.962	0.091
`qwen2.5-7b-instruct`	7.6B	0.207	0.203	0.168	0.172	0.964	0.083
`deepseek-llm-7b-chat`	7B	0.207	0.204	0.170	0.174	0.940	0.128
`qwen3-4b-instruct`	4B	0.197	0.193	0.159	0.163	0.958	0.094
LLMs (base)							
`llama3-8b`	8B	**0.254**	**0.250**	**0.210**	**0.213**	**0.967**	**0.079**
`llama3.1-8b`	8B	0.253	0.249	0.209	0.212	0.967	0.080
`deepseek-llm-7b-base`	7B	0.220	0.216	0.180	0.184	0.954	0.105
`qwen2.5-3b`	3.1B	0.212	0.208	0.172	0.176	0.962	0.090
`qwen3-4b`	4B	0.210	0.206	0.171	0.175	0.960	0.087
Embedding models (instruction-tuned)							
`e5-mistral-7b-instruct`	7.1B	**0.242**	**0.238**	**0.202**	**0.205**	0.955	0.096
`gte-qwen2-7b-instruct`	7.6B	0.240	0.235	0.198	0.202	**0.961**	**0.078**
Embedding models							
`multilingual-e5-large`	559.9M	**0.258**	**0.254**	**0.217**	**0.220**	0.920	0.157
`multilingual-e5-base`	278M	0.250	0.245	0.209	0.212	0.918	0.161
`qwen3-embedding-8b`	7.6B	0.241	0.237	0.201	0.203	0.940	0.135
`sfr-embedding-mistral`	7.1B	0.240	0.236	0.201	0.205	**0.956**	**0.096**
`qwen3-embedding-4b`	4B	0.236	0.231	0.196	0.198	0.953	0.111
`snowflake-arctic-embed-l-v2`	567.8M	0.212	0.208	0.182	0.185	0.864	0.219
`qwen3-embedding-0.6b`	595.8M	0.211	0.208	0.177	0.181	0.953	0.108
`e5-large-v2`	335.1M	0.208	0.205	0.173	0.176	0.887	0.187
`e5-base-v2`	109M	0.200	0.197	0.165	0.169	0.888	0.179
`facebook-contriever`	110M	0.195	0.192	0.160	0.164	0.895	0.169
`gte-large-en-v1.5`	409M	0.189	0.187	0.156	0.159	0.923	0.129
`bge-m3`	567M	0.188	0.185	0.161	0.164	0.796	0.281
`bge-large-en-v1.5`	335M	0.177	0.175	0.147	0.150	0.867	0.192
`e5-small-v2`	33M	0.177	0.174	0.146	0.150	0.856	0.216
`gte-large`	335.1M	0.175	0.172	0.144	0.148	0.889	0.163
`mxbai-embed-large-v1`	335.1M	0.174	0.171	0.143	0.146	0.865	0.197
`bge-base-en-v1.5`	109.5M	0.172	0.169	0.142	0.145	0.850	0.211
`gte-base`	109.5M	0.168	0.165	0.138	0.141	0.877	0.178
`bge-base-zh-v1.5`	102M	0.167	0.164	0.139	0.143	0.912	0.152
`facebook-contriever-msmarco`	110M	0.167	0.164	0.138	0.142	0.870	0.200
`bge-large-zh-v1.5`	326M	0.163	0.159	0.135	0.139	0.903	0.169
`distiluse-base-multilingual-cased-v2`	134.7M	0.162	0.159	0.134	0.137	0.836	0.245
`all-roberta-large-v1`	355.4M	0.161	0.158	0.132	0.136	0.908	0.144
`all-mpnet-base-v2`	109.5M	0.160	0.157	0.130	0.133	0.889	0.169
`bge-small-en-v1.5`	33.4M	0.158	0.155	0.131	0.135	0.853	0.214
`all-minilm-l12-v2`	33.4M	0.152	0.150	0.125	0.129	0.888	0.185
`paraphrase-mpnet-base-v2`	109.5M	0.149	0.146	0.123	0.127	0.881	0.181
`bert-base-uncased`	110.1M	0.145	0.143	0.118	0.122	0.944	0.101
`all-minilm-l6-v2`	22.7M	0.145	0.143	0.120	0.124	0.891	0.176
`paraphrase-multilingual-mpnet-base-v2`	278M	0.138	0.136	0.117	0.119	0.794	0.281
`msmarco-distilbert-base-v4`	66.4M	0.133	0.130	0.110	0.114	0.838	0.247
`allenai-specter`	110M	0.107	0.105	0.086	0.094	0.901	0.169
`jina-embeddings-v2-small-en`	33M	0.100	0.098	0.081	0.086	0.935	0.121
`jina-embeddings-v2-base-en`	137.4M	0.016	0.016	0.012	0.014	0.738	0.328
Lexical / non-neural baselines							
`tfidf`	–	**0.197**	**0.194**	**0.167**	**0.183**	0.837	0.219
`ngram`	–	0.170	0.167	0.145	0.149	**0.874**	**0.213**
`ppm`	–	0.157	0.156	0.137	0.142	0.792	0.298

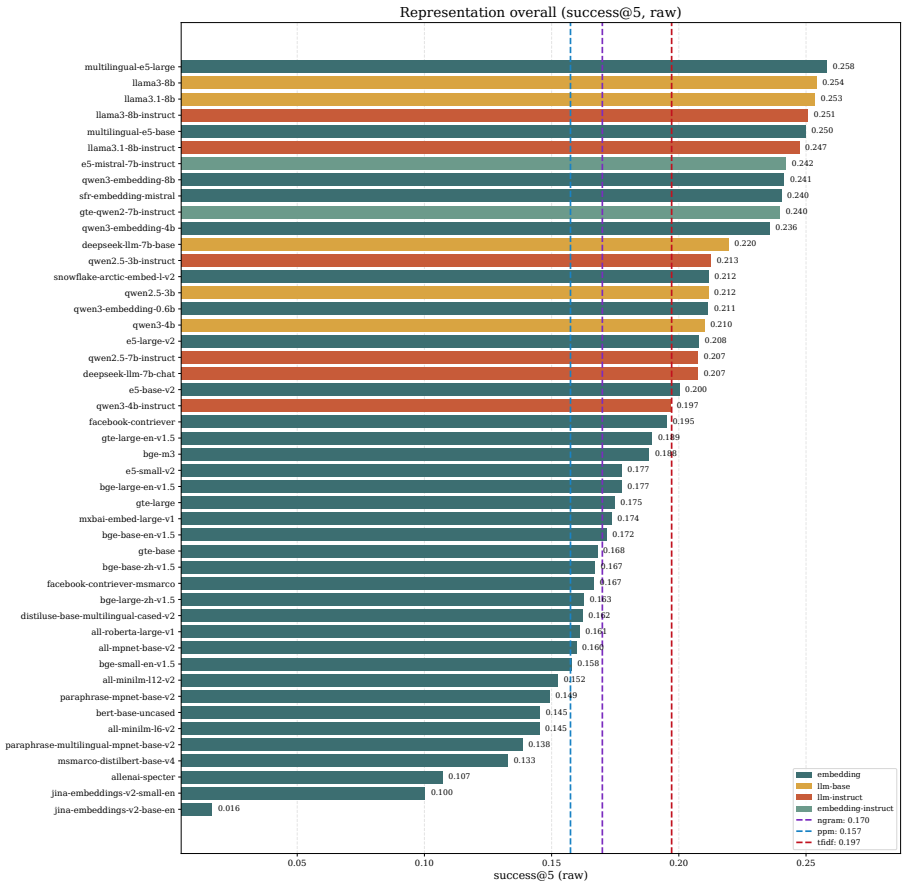

Figure 4: Overall AuthBench performance by Success@5 (S@5).

B.3 Language-wise Full Results

The following tables provide the complete language-level breakdown for both retrieval and verification-oriented evaluation.

B.4 Primary-genre Full Results

These tables report the full genre-level breakdown, highlighting how strongly authorship performance depends on discourse type.

B.5 Length-bucket Full Results

These tables give the complete length-wise breakdown and make the sparse-text versus long-document gap explicit across all evaluated systems.

C Post-Training Note

This appendix focuses on the zero-shot benchmark results reported in the current paper. In parallel, we are conducting a broader post-training study covering multiple adaptation strategies and authorship methods, and we plan to present those findings in a second version of the paper. The zero-shot results here are intended to provide a clean foundation for that next stage of analysis.

D Additional Dataset Statistics

This appendix complements the full result tables with additional dataset-profile visualizations. Figures 4–9 summarize overall metric-level performance across models, Figure 10 shows token-length distributions by language, and Figure 11 shows the fine-grained subgenre mix within each language.

E Raw Data Sources

AuthBench draws on 17 publicly available input sources spanning multiple platforms, domains, and languages. Each raw item is mapped into a unified schema (Section 3.1) with provenance (`source_id`), language (`lang`), a normalized genre label (`genre`), and token length $\ell(d) \in \mathbb{N}$ computed under a fixed tokenizer. Table 15 lists these sources, and Table 16 reports their realized contribution to the current benchmark release.

Table 9: Language-wise Success@5 on AuthBench (full results).

Model	Model Size	ar	de	en	es	fr	hi	ja	ko	ru	zh
LLMs (instruction-tuned)											
deepseek-llm-7b-chat	7B	0.170	0.203	0.223	0.223	0.213	0.204	0.184	0.135	0.126	0.347
llama3-8b-instruct	8B	**0.226**	0.244	**0.261**	**0.265**	**0.269**	0.289	**0.243**	0.173	0.164	**0.382**
llama3.1-8b-instruct	8B	0.211	**0.245**	0.259	0.265	0.265	**0.297**	0.229	**0.177**	**0.165**	0.373
qwen2.5-3b-instruct	3.1B	0.207	0.206	0.206	0.215	0.220	0.218	0.223	0.145	0.150	0.336
qwen2.5-7b-instruct	7.6B	0.182	0.210	0.201	0.208	0.214	0.215	0.215	0.143	0.143	0.343
qwen3-4b-instruct	4B	0.187	0.195	0.187	0.203	0.198	0.229	0.212	0.143	0.143	0.304
LLMs (base)											
deepseek-llm-7b-base	7B	0.183	0.216	0.238	0.223	0.220	0.238	0.200	0.144	0.131	0.367
llama3-8b	8B	0.221	0.247	**0.270**	**0.272**	**0.270**	0.289	0.242	0.183	0.165	**0.380**
llama3.1-8b	8B	**0.225**	**0.249**	0.270	0.272	0.266	**0.297**	**0.246**	**0.186**	**0.173**	0.361
qwen2.5-3b	3.1B	0.206	0.209	0.209	0.210	0.208	0.215	0.221	0.147	0.148	0.333
qwen3-4b	4B	0.189	0.217	0.204	0.213	0.214	0.227	0.220	0.142	0.147	0.336
Embedding models (instruction-tuned)											
e5-mistral-7b-instruct	7.1B	**0.223**	**0.241**	**0.237**	**0.254**	**0.250**	**0.272**	0.223	**0.164**	0.164	**0.396**
gte-qwen2-7b-instruct	7.6B	0.220	0.241	0.232	0.235	0.239	0.252	**0.249**	0.156	**0.171**	0.396
Embedding models											
all-minilm-l12-v2	33.4M	0.128	0.135	0.181	0.157	0.162	0.195	0.135	0.083	0.084	0.246
all-minilm-l6-v2	22.7M	0.117	0.128	0.173	0.157	0.152	0.184	0.116	0.076	0.077	0.245
all-mpnet-base-v2	109.5M	0.152	0.154	0.190	0.166	0.166	0.176	0.148	0.092	0.092	0.230
all-roberta-large-v1	355.4M	0.159	0.180	0.195	0.195	0.181	0.198	0.155	0.091	0.094	0.170
allenai-specter	110M	0.092	0.139	0.133	0.126	0.133	0.133	0.055	0.080	0.066	0.101
bert-base-uncased	110.1M	0.135	0.136	0.188	0.140	0.155	0.156	0.112	0.095	0.079	0.199
bge-base-en-v1.5	109.5M	0.155	0.171	0.193	0.200	0.203	0.195	0.148	0.099	0.101	0.241
bge-base-zh-v1.5	102M	0.116	0.124	0.147	0.153	0.137	0.193	0.190	0.124	0.086	0.407
bge-large-en-v1.5	335M	0.153	0.176	0.198	0.205	0.207	0.184	0.157	0.105	0.114	0.249
bge-large-zh-v1.5	326M	0.113	0.127	0.144	0.147	0.131	0.176	0.175	0.116	0.085	0.401
bge-m3	567M	0.213	0.159	0.155	0.174	0.188	0.176	0.165	0.138	0.129	0.369
bge-small-en-v1.5	33.4M	0.131	0.145	0.184	0.180	0.175	0.139	0.132	0.089	0.087	0.253
distiluse-base-multilingual-cased-v2	134.7M	0.178	0.148	0.160	0.148	0.142	0.193	0.139	0.109	0.111	0.281
e5-base-v2	109M	0.182	0.201	0.227	0.242	0.236	0.215	0.194	0.110	0.128	0.255
e5-large-v2	335.1M	0.203	0.197	0.229	0.248	0.249	0.210	0.192	0.120	0.141	0.269
e5-small-v2	33M	0.150	0.178	0.201	0.199	0.193	0.207	0.171	0.092	0.116	0.254
facebook-contriever	110M	0.172	0.187	**0.261**	0.228	0.236	0.246	0.158	0.116	0.117	0.200
facebook-contriever-msmarco	110M	0.152	0.151	0.208	0.197	0.178	0.201	0.130	0.099	0.094	0.224
gte-base	109.5M	0.119	0.176	0.210	0.206	0.205	0.125	0.149	0.080	0.099	0.224
gte-large	335.1M	0.113	0.182	0.216	0.219	0.217	0.139	0.149	0.084	0.113	0.232
gte-large-en-v1.5	409M	0.181	0.185	0.223	0.228	0.222	0.210	0.180	0.111	0.115	0.231
jina-embeddings-v2-base-en	137.4M	0.050	0.008	0.005	0.005	0.006	0.065	0.029	0.013	0.013	0.027
jina-embeddings-v2-small-en	33M	0.082	0.077	0.108	0.101	0.081	0.119	0.097	0.081	0.057	0.188
msmarco-distilbert-base-v4	66.4M	0.123	0.113	0.157	0.127	0.115	0.142	0.102	0.091	0.079	0.224
multilingual-e5-base	278M	0.251	0.210	0.226	0.267	0.262	0.238	0.253	0.185	0.188	0.413
multilingual-e5-large	559.9M	**0.252**	0.221	0.231	**0.284**	**0.280**	**0.263**	**0.268**	**0.196**	**0.190**	0.422
mxbai-embed-large-v1	335.1M	0.143	0.173	0.196	0.198	0.204	0.193	0.154	0.101	0.116	0.240
paraphrase-mpnet-base-v2	109.5M	0.134	0.136	0.182	0.163	0.138	0.210	0.139	0.098	0.091	0.203
paraphrase-multilingual-mpnet-base-v2	278M	0.151	0.111	0.124	0.127	0.123	0.147	0.123	0.091	0.102	0.269
qwen3-embedding-0.6b	595.8M	0.208	0.192	0.188	0.207	0.200	0.221	0.219	0.147	0.139	0.399
qwen3-embedding-4b	4B	0.220	0.231	0.213	0.227	0.243	0.218	0.241	0.167	0.160	0.422
qwen3-embedding-8b	7.6B	0.239	**0.242**	0.209	0.239	0.241	0.244	0.251	0.169	0.170	**0.425**
sfr-embedding-mistral	7.1B	0.221	0.236	0.236	0.255	0.248	0.258	0.214	0.162	0.162	0.397
snowflake-arctic-embed-l-v2	567.8M	0.240	0.161	0.184	0.187	0.202	0.252	0.215	0.154	0.160	0.385
Lexical / non-neural baselines											
ngram	–	0.238	0.137	0.126	0.162	0.166	0.244	0.160	0.168	0.132	0.267
ppm	–	0.229	0.140	**0.149**	0.190	**0.172**	0.255	0.113	0.103	0.118	0.184
tfidf	–	**0.254**	**0.148**	0.146	**0.190**	0.155	**0.280**	**0.226**	**0.184**	**0.154**	**0.346**

Table 10: Language-wise EER on AuthBench (full results).

Model	Model Size	ar	de	en	es	fr	hi	ja	ko	ru	zh
LLMs (instruction-tuned)											
deepseek-llm-7b-chat	7B	0.125	0.103	0.141	0.089	0.103	0.071	0.061	0.081	0.123	0.074
llama3-8b-instruct	8B	0.073	0.064	0.099	**0.064**	**0.061**	0.051	**0.042**	**0.054**	**0.112**	0.057
llama3.1-8b-instruct	8B	0.073	**0.062**	**0.096**	0.066	0.062	**0.045**	0.042	0.055	0.114	0.059
qwen2.5-3b-instruct	3.1B	**0.072**	0.072	0.131	0.074	0.068	0.054	0.043	0.060	0.119	0.062
qwen2.5-7b-instruct	7.6B	0.075	0.070	0.118	0.068	0.064	0.048	0.044	0.059	0.120	**0.056**
qwen3-4b-instruct	4B	0.080	0.071	0.131	0.084	0.078	0.057	0.046	0.063	0.121	0.061
LLMs (base)											
deepseek-llm-7b-base	7B	0.082	0.072	0.128	0.076	0.071	0.062	0.043	0.061	0.121	0.065
llama3-8b	8B	0.074	**0.062**	**0.094**	0.064	0.061	0.051	**0.041**	**0.052**	0.110	**0.056**
llama3.1-8b	8B	**0.073**	0.062	0.094	**0.064**	**0.060**	**0.045**	0.042	0.053	**0.109**	0.060
qwen2.5-3b	3.1B	0.074	0.072	0.129	0.074	0.069	0.057	0.043	0.060	0.119	0.060
qwen3-4b	4B	0.079	0.066	0.123	0.077	0.068	0.057	0.045	0.064	0.123	0.059
Embedding models (instruction-tuned)											
e5-mistral-7b-instruct	7.1B	**0.072**	0.072	0.113	**0.067**	0.064	0.048	0.043	0.063	**0.118**	0.061
gte-qwen2-7b-instruct	7.6B	0.073	**0.069**	**0.107**	0.067	**0.063**	**0.045**	**0.041**	**0.061**	0.118	**0.054**
Embedding models											
all-minilm-l12-v2	33.4M	0.085	0.207	0.219	0.107	0.123	0.062	0.120	0.105	0.122	0.102
all-minilm-l6-v2	22.7M	0.088	0.156	0.217	0.099	0.109	0.054	0.097	0.099	**0.114**	0.103
all-mpnet-base-v2	109.5M	0.081	0.132	0.226	0.096	0.101	0.051	0.095	0.087	0.115	0.107
all-roberta-large-v1	355.4M	0.082	0.083	0.192	0.072	0.082	0.042	0.082	0.065	0.117	0.096
allenai-specter	110M	0.188	0.147	0.208	0.167	0.168	0.210	0.181	0.170	0.132	0.164
bert-base-uncased	110.1M	0.084	0.074	0.151	0.078	0.076	0.054	0.070	0.070	0.124	0.106
bge-base-en-v1.5	109.5M	0.082	0.140	0.309	0.144	0.135	0.049	0.090	0.074	0.129	0.129
bge-base-zh-v1.5	102M	0.103	0.131	0.190	0.116	0.152	0.096	0.063	0.070	0.124	0.172
bge-large-en-v1.5	335M	0.079	0.115	0.287	0.122	0.133	0.049	0.078	0.077	0.132	0.119
bge-large-zh-v1.5	326M	0.094	0.173	0.241	0.164	0.192	0.126	0.104	0.099	0.124	0.205
bge-m3	567M	0.265	0.273	0.286	0.302	0.265	0.320	0.275	0.280	0.330	0.216
bge-small-en-v1.5	33.4M	0.099	0.130	0.250	0.159	0.128	0.068	0.094	0.116	0.144	0.140
distiluse-base-multilingual-cased-v2	134.7M	0.253	0.256	0.206	0.262	0.241	0.271	0.235	0.238	0.308	0.210
e5-base-v2	109M	0.077	0.118	0.242	0.116	0.122	0.071	0.097	0.082	0.132	0.107
e5-large-v2	335.1M	0.099	0.159	0.244	0.148	0.136	0.068	0.139	0.092	0.169	0.139
e5-small-v2	33M	0.100	0.189	0.278	0.185	0.177	0.092	0.149	0.086	0.167	0.130
facebook-contriever	110M	0.076	0.096	0.177	0.095	0.095	0.051	0.052	0.067	0.118	0.084
facebook-contriever-msmarco	110M	0.086	0.143	0.228	0.127	0.170	0.068	0.067	0.072	0.125	0.106
gte-base	109.5M	0.079	0.093	0.264	0.097	0.092	0.042	0.068	0.074	0.127	0.113
gte-large	335.1M	0.079	0.090	0.263	0.091	0.086	**0.037**	0.061	0.075	0.124	0.113
gte-large-en-v1.5	409M	0.081	0.081	0.216	0.073	0.068	0.040	0.078	0.109	0.126	0.123
jina-embeddings-v2-base-en	137.4M	0.201	0.345	0.376	0.349	0.352	0.178	0.289	0.350	0.287	0.252
jina-embeddings-v2-small-en	33M	0.081	0.142	0.164	0.136	0.144	0.057	0.078	0.071	0.119	0.094
msmarco-distilbert-base-v4	66.4M	0.083	0.118	0.249	0.111	0.122	0.076	0.115	0.083	0.131	0.119
multilingual-e5-base	278M	0.140	0.143	0.201	0.170	0.137	0.213	0.096	0.098	0.185	0.078
multilingual-e5-large	559.9M	0.133	0.136	0.202	0.147	0.124	0.190	0.089	0.106	0.176	0.073
mxbai-embed-large-v1	335.1M	0.080	0.124	0.285	0.130	0.139	0.051	0.073	0.074	0.127	0.124
paraphrase-mpnet-base-v2	109.5M	0.081	0.103	0.221	0.087	0.088	0.099	0.104	0.088	0.127	0.121
paraphrase-multilingual-mpnet-base-v2	278M	0.225	0.274	0.298	0.301	0.278	0.278	0.249	0.257	0.296	0.249
qwen3-embedding-0.6b	595.8M	0.115	0.103	0.148	0.111	0.072	0.127	0.049	0.065	0.123	0.087
qwen3-embedding-4b	4B	0.092	0.112	0.156	0.115	0.091	0.089	0.051	0.063	0.122	0.092
qwen3-embedding-8b	7.6B	**0.066**	0.135	0.174	0.148	0.114	0.051	0.045	**0.059**	0.114	0.074
sfr-embedding-mistral	7.1B	0.071	**0.073**	**0.112**	**0.067**	**0.064**	0.048	**0.043**	0.063	0.117	**0.061**
snowflake-arctic-embed-l-v2	567.8M	0.167	0.244	0.233	0.248	0.225	0.164	0.198	0.200	0.255	0.172
Lexical / non-neural baselines											
ngram	–	0.220	0.191	0.213	0.194	0.194	0.110	**0.134**	**0.257**	0.262	**0.119**
ppm	–	0.288	0.262	0.257	0.257	0.248	0.278	0.347	0.344	0.359	0.323
tfidf	–	**0.136**	**0.165**	**0.200**	**0.179**	**0.188**	**0.082**	0.347	0.340	**0.145**	0.354

Table 11: Primary-genre Success@5 on AuthBench (full results).

Model	Model Size	blog	ecomm	literature	media	news	poetry	qna	research	social
LLMs (instruction-tuned)										
deepseek-llm-7b-chat	7B	0.196	0.110	0.461	0.053	0.243	0.556	0.541	0.713	0.166
llama3-8b-instruct	8B	0.213	0.107	**0.521**	**0.070**	**0.302**	**0.639**	**0.650**	0.769	**0.206**
llama3.1-8b-instruct	8B	**0.220**	**0.112**	0.518	0.035	0.293	0.583	0.618	**0.787**	0.204
qwen2.5-3b-instruct	3.1B	0.171	0.100	0.483	0.035	0.252	0.639	0.573	0.731	0.171
qwen2.5-7b-instruct	7.6B	0.162	0.084	0.496	0.035	0.242	0.583	0.567	0.719	0.167
qwen3-4b-instruct	4B	0.149	0.080	0.462	0.053	0.227	0.611	0.548	0.688	0.160
LLMs (base)										
deepseek-llm-7b-base	7B	0.215	0.114	0.486	0.035	0.257	0.639	0.548	0.719	0.176
llama3-8b	8B	0.229	**0.119**	0.524	**0.070**	**0.311**	**0.667**	**0.631**	0.800	**0.207**
llama3.1-8b	8B	**0.232**	0.112	**0.532**	0.053	0.305	0.667	0.631	**0.812**	0.207
qwen2.5-3b	3.1B	0.178	0.096	0.487	0.018	0.248	0.639	0.573	0.744	0.170
qwen3-4b	4B	0.166	0.084	0.485	0.035	0.241	0.639	0.561	0.738	0.171
Embedding models (instruction-tuned)										
e5-mistral-7b-instruct	7.1B	**0.202**	**0.110**	0.514	**0.053**	0.262	0.583	0.580	0.725	**0.207**
gte-qwen2-7b-instruct	7.6B	0.191	0.091	**0.538**	0.035	**0.281**	**0.611**	**0.586**	**0.781**	0.197
Embedding models										
all-minilm-l12-v2	33.4M	0.128	0.066	0.265	0.018	0.138	0.222	0.338	0.781	0.141
all-minilm-l6-v2	22.7M	0.120	0.055	0.241	0.018	0.129	0.278	0.338	0.756	0.136
all-mpnet-base-v2	109.5M	0.131	0.082	0.301	0.018	0.144	0.250	0.465	0.838	0.144
all-roberta-large-v1	355.4M	0.132	0.078	0.285	0.018	0.156	0.250	0.408	0.756	0.146
allenai-specter	110M	0.093	0.059	0.265	0.018	0.097	0.222	0.255	0.644	0.089
bert-base-uncased	110.1M	0.154	0.082	0.311	0.035	0.146	0.222	0.280	0.569	0.124
bge-base-en-v1.5	109.5M	0.124	0.082	0.318	0.018	0.162	0.250	0.465	**0.844**	0.156
bge-base-zh-v1.5	102M	0.082	0.068	0.310	0.070	0.173	0.250	0.382	0.463	0.155
bge-large-en-v1.5	335M	0.143	0.073	0.342	0.018	0.161	0.306	0.522	0.838	0.161
bge-large-zh-v1.5	326M	0.082	0.073	0.305	0.035	0.167	0.250	0.408	0.431	0.151
bge-m3	567M	0.108	0.066	0.419	0.018	0.146	0.389	0.433	0.594	0.181
bge-small-en-v1.5	33.4M	0.116	0.075	0.273	0.018	0.140	0.194	0.490	0.800	0.147
distiluse-base-multilingual-cased-v2	134.7M	0.104	0.080	0.326	0.000	0.149	0.472	0.497	0.600	0.148
e5-base-v2	109M	0.151	0.098	0.369	0.018	0.212	0.333	0.592	0.800	0.176
e5-large-v2	335.1M	0.155	0.089	0.388	0.018	0.216	0.306	0.643	0.819	0.184
e5-small-v2	33M	0.124	0.098	0.338	0.035	0.182	0.306	0.567	0.750	0.156
facebook-contriever	110M	0.193	**0.114**	0.380	0.018	0.237	0.361	0.433	0.744	0.159
facebook-contriever-msmarco	110M	0.154	0.094	0.292	0.018	0.171	0.250	0.439	0.738	0.147
gte-base	109.5M	0.143	0.084	0.320	0.018	0.167	0.194	0.522	0.831	0.146
gte-large	335.1M	0.144	0.068	0.343	0.018	0.171	0.278	0.529	0.831	0.154
gte-large-en-v1.5	409M	0.141	0.082	0.351	0.018	0.208	0.278	0.561	0.806	0.165
jina-embeddings-v2-base-en	137.4M	0.005	0.002	0.028	0.000	0.012	0.000	0.006	0.006	0.018
jina-embeddings-v2-small-en	33M	0.063	0.041	0.194	0.018	0.089	0.083	0.127	0.338	0.096
msmarco-distilbert-base-v4	66.4M	0.108	0.057	0.225	0.035	0.135	0.167	0.293	0.625	0.120
multilingual-e5-base	278M	0.146	0.107	0.453	0.053	0.258	0.500	0.631	0.756	0.230
multilingual-e5-large	559.9M	0.160	0.100	0.479	0.053	**0.268**	0.444	**0.682**	0.769	**0.236**
mxbai-embed-large-v1	335.1M	0.134	0.064	0.345	0.018	0.158	0.167	0.522	0.831	0.157
paraphrase-mpnet-base-v2	109.5M	0.146	0.094	0.274	0.018	0.140	0.278	0.414	0.694	0.132
paraphrase-multilingual-mpnet-base-v2	278M	0.081	0.053	0.295	0.000	0.101	0.389	0.433	0.644	0.132
qwen3-embedding-0.6b	595.8M	0.124	0.078	0.461	0.053	0.211	0.556	0.529	0.787	0.188
qwen3-embedding-4b	4B	0.151	0.084	0.535	0.070	0.243	**0.583**	0.561	0.831	0.205
qwen3-embedding-8b	7.6B	0.151	0.091	**0.536**	**0.088**	0.256	0.556	0.586	0.838	0.209
sfr-embedding-mistral	7.1B	**0.199**	0.105	0.512	0.053	0.257	0.583	0.567	0.756	0.207
snowflake-arctic-embed-l-v2	567.8M	0.129	0.057	0.418	0.035	0.178	0.472	0.541	0.738	0.202
Lexical / non-neural baselines										
ngram	–	0.066	0.050	0.289	0.018	0.161	0.278	0.325	0.431	0.168
ppm	–	**0.090**	**0.064**	0.230	0.000	0.149	0.250	0.420	**0.631**	0.153
tfidf	–	0.089	0.053	**0.323**	**0.035**	**0.169**	**0.444**	**0.459**	0.613	**0.198**

Table 12: Primary-genre EER on AuthBench (full results).

Model	Model Size	blog	ecomm	literature	media	news	poetry	qna	research	social
LLMs (instruction-tuned)										
deepseek-llm-7b-chat	7B	0.133	0.094	0.076	0.078	0.111	0.090	0.110	0.029	0.132
llama3-8b-instruct	8B	**0.106**	0.074	**0.036**	0.078	0.070	0.064	**0.063**	**0.011**	0.081
llama3.1-8b-instruct	8B	0.108	**0.064**	0.037	0.080	**0.065**	0.068	0.067	0.012	**0.081**
qwen2.5-3b-instruct	3.1B	0.137	0.086	0.048	0.079	0.075	**0.059**	0.098	0.031	0.091
qwen2.5-7b-instruct	7.6B	0.131	0.080	0.051	0.075	0.070	0.061	0.087	0.019	0.085
qwen3-4b-instruct	4B	0.136	0.096	0.050	**0.073**	0.081	0.080	0.098	0.046	0.092
LLMs (base)										
deepseek-llm-7b-base	7B	0.135	0.091	0.070	0.078	0.103	0.079	0.099	0.031	0.105
llama3-8b	8B	**0.105**	**0.062**	0.036	**0.073**	**0.070**	0.068	0.069	**0.009**	**0.081**
llama3.1-8b	8B	0.108	0.063	**0.035**	0.081	0.071	0.072	**0.067**	0.009	0.082
qwen2.5-3b	3.1B	0.134	0.085	0.049	0.077	0.074	**0.065**	0.101	0.031	0.090
qwen3-4b	4B	0.128	0.084	0.046	0.074	0.074	0.078	0.091	0.025	0.088
Embedding models (instruction-tuned)										
e5-mistral-7b-instruct	7.1B	0.120	0.073	0.055	0.085	0.088	0.078	0.090	0.025	0.093
gte-qwen2-7b-instruct	7.6B	**0.113**	**0.070**	**0.047**	**0.069**	**0.067**	**0.076**	**0.075**	**0.013**	**0.082**
Embedding models										
all-minilm-l12-v2	33.4M	0.214	0.267	0.111	0.138	0.194	0.063	0.126	0.025	0.167
all-minilm-l6-v2	22.7M	0.211	0.263	0.084	0.133	0.200	0.081	0.109	0.025	0.146
all-mpnet-base-v2	109.5M	0.219	0.277	0.079	0.138	0.197	0.059	0.121	0.019	0.138
all-roberta-large-v1	355.4M	0.206	0.105	0.076	0.108	0.175	**0.039**	0.121	0.031	0.114
allenai-specter	110M	0.164	0.172	0.070	0.192	0.164	0.078	0.098	0.031	0.182
bert-base-uncased	110.1M	0.128	0.126	0.069	0.137	0.102	0.082	0.082	0.029	0.101
bge-base-en-v1.5	109.5M	0.230	0.365	0.078	0.211	0.298	0.082	0.136	0.031	0.175
bge-base-zh-v1.5	102M	0.157	0.142	0.061	0.310	0.142	0.062	0.080	0.056	0.152
bge-large-en-v1.5	335M	0.222	0.322	0.080	0.183	0.259	0.098	0.117	**0.013**	0.151
bge-large-zh-v1.5	326M	0.211	0.202	0.066	0.336	0.154	0.063	0.086	0.102	0.162
bge-m3	567M	0.257	0.214	0.076	0.368	0.265	0.118	0.109	0.054	0.304
bge-small-en-v1.5	33.4M	0.202	0.277	0.099	0.229	0.264	0.092	0.148	0.025	0.178
distiluse-base-multilingual-cased-v2	134.7M	0.191	0.145	0.091	0.391	0.191	0.122	0.052	0.046	0.277
e5-base-v2	109M	0.215	0.194	0.072	0.172	0.217	0.056	0.057	0.025	0.162
e5-large-v2	335.1M	0.217	0.180	0.078	0.207	0.221	0.078	0.059	0.031	0.179
e5-small-v2	33M	0.244	0.203	0.095	0.230	0.264	0.063	0.063	0.050	0.205
facebook-contriever	110M	0.228	0.094	0.075	0.111	0.155	0.081	0.109	0.013	0.142
facebook-contriever-msmarco	110M	0.234	0.148	0.090	0.127	0.211	0.078	0.095	0.021	0.179
gte-base	109.5M	0.210	0.359	0.077	0.155	0.243	0.085	0.109	0.031	0.138
gte-large	335.1M	0.207	0.358	0.066	0.153	0.213	0.084	0.098	0.025	0.127
gte-large-en-v1.5	409M	0.193	0.516	0.042	0.149	0.122	0.059	0.079	0.031	0.117
jina-embeddings-v2-base-en	137.4M	0.326	0.352	0.275	0.230	0.342	0.255	0.337	0.294	0.325
jina-embeddings-v2-small-en	33M	0.159	0.126	0.083	0.125	0.124	0.083	0.094	0.082	0.122
msmarco-distilbert-base-v4	66.4M	0.238	0.205	0.117	0.146	0.241	0.089	0.208	0.025	0.234
multilingual-e5-base	278M	0.179	0.132	0.087	0.090	0.145	0.137	**0.027**	0.025	0.151
multilingual-e5-large	559.9M	0.197	0.160	0.078	0.105	0.146	0.118	0.034	0.022	0.147
mxbai-embed-large-v1	335.1M	0.242	0.311	0.081	0.175	0.253	0.096	0.116	0.013	0.156
paraphrase-mpnet-base-v2	109.5M	0.218	0.128	0.079	0.144	0.228	0.093	0.114	0.019	0.142
paraphrase-multilingual-mpnet-base-v2	278M	0.310	0.212	0.122	0.368	0.287	0.176	0.098	0.025	0.283
qwen3-embedding-0.6b	595.8M	0.137	0.096	0.044	0.166	0.103	0.066	0.063	0.015	0.109
qwen3-embedding-4b	4B	0.131	0.095	**0.040**	0.174	0.110	0.061	0.080	0.013	0.107
qwen3-embedding-8b	7.6B	0.166	0.097	0.070	0.120	0.139	0.070	0.092	0.025	0.126
sfr-embedding-mistral	7.1B	**0.117**	**0.074**	0.053	**0.088**	**0.089**	0.088	0.088	0.025	**0.093**
snowflake-arctic-embed-l-v2	567.8M	0.220	0.158	0.065	0.324	0.210	0.078	0.128	0.056	0.233
Lexical / non-neural baselines										
ngram	–	0.195	0.234	**0.128**	**0.134**	**0.146**	0.127	0.133	0.087	**0.231**
ppm	–	0.214	0.240	0.275	0.356	0.276	0.335	0.207	**0.069**	0.315
tfidf	–	**0.154**	**0.185**	0.148	0.498	0.199	**0.098**	**0.079**	0.075	0.243

Table 13: Length-bucket Success@5 on AuthBench (full results).

Model	Model Size	short	medium	long	extra_long
LLMs (instruction-tuned)					
deepseek-llm-7b-chat	7B	0.106	0.161	0.371	0.524
llama3-8b-instruct	8B	**0.150**	**0.199**	**0.432**	**0.539**
llama3.1-8b-instruct	8B	0.142	0.196	0.427	0.531
qwen2.5-3b-instruct	3.1B	0.101	0.164	0.385	0.508
qwen2.5-7b-instruct	7.6B	0.103	0.159	0.379	0.496
qwen3-4b-instruct	4B	0.081	0.150	0.368	0.488
LLMs (base)					
deepseek-llm-7b-base	7B	0.121	0.169	0.393	0.528
llama3-8b	8B	**0.150**	**0.202**	0.436	0.524
llama3.1-8b	8B	0.135	0.202	**0.439**	**0.543**
qwen2.5-3b	3.1B	0.093	0.163	0.386	0.520
qwen3-4b	4B	0.101	0.161	0.385	0.496
Embedding models (instruction-tuned)					
e5-mistral-7b-instruct	7.1B	**0.161**	**0.196**	0.398	**0.524**
gte-qwen2-7b-instruct	7.6B	0.137	0.192	**0.409**	0.500
Embedding models					
all-minilm-l12-v2	33.4M	0.128	0.129	0.222	0.343
all-minilm-l6-v2	22.7M	0.126	0.122	0.211	0.335
all-mpnet-base-v2	109.5M	0.122	0.133	0.246	0.358
all-roberta-large-v1	355.4M	0.113	0.136	0.248	0.280
allenai-specter	110M	0.069	0.076	0.206	0.299
bert-base-uncased	110.1M	0.095	0.113	0.250	0.354
bge-base-en-v1.5	109.5M	0.117	0.145	0.265	0.323
bge-base-zh-v1.5	102M	0.166	0.132	0.263	0.386
bge-large-en-v1.5	335M	0.124	0.149	0.274	0.346
bge-large-zh-v1.5	326M	0.164	0.130	0.251	0.370
bge-m3	567M	0.151	0.152	0.302	0.406
bge-small-en-v1.5	33.4M	0.125	0.135	0.231	0.319
distiluse-base-multilingual-cased-v2	134.7M	0.114	0.130	0.270	0.327
e5-base-v2	109M	0.137	0.169	0.307	0.390
e5-large-v2	335.1M	0.140	0.176	0.320	0.406
e5-small-v2	33M	0.123	0.146	0.282	0.378
facebook-contriever	110M	0.120	0.161	0.317	0.402
facebook-contriever-msmarco	110M	0.116	0.140	0.257	0.366
gte-base	109.5M	0.112	0.144	0.255	0.323
gte-large	335.1M	0.118	0.149	0.262	0.370
gte-large-en-v1.5	409M	0.108	0.162	0.293	0.406
jina-embeddings-v2-base-en	137.4M	0.011	0.013	0.026	0.067
jina-embeddings-v2-small-en	33M	0.088	0.078	0.161	0.299
msmarco-distilbert-base-v4	66.4M	0.115	0.109	0.204	0.291
multilingual-e5-base	278M	**0.187**	0.213	0.376	0.417
multilingual-e5-large	559.9M	0.182	**0.220**	0.392	0.476
mxbai-embed-large-v1	335.1M	0.116	0.147	0.267	0.346
paraphrase-mpnet-base-v2	109.5M	0.113	0.119	0.244	0.335
paraphrase-multilingual-mpnet-base-v2	278M	0.097	0.111	0.229	0.307
qwen3-embedding-0.6b	595.8M	0.154	0.173	0.337	0.453
qwen3-embedding-4b	4B	0.167	0.191	0.384	0.500
qwen3-embedding-8b	7.6B	0.157	0.197	**0.395**	0.504
sfr-embedding-mistral	7.1B	0.168	0.193	0.395	**0.520**
snowflake-arctic-embed-l-v2	567.8M	0.161	0.175	0.329	0.445
Lexical / non-neural baselines					
ngram	–	0.151	0.136	0.271	0.331
ppm	–	0.079	0.134	0.255	0.240
tfidf	–	**0.157**	**0.164**	**0.300**	**0.437**

Table 14: Length-bucket EER on AuthBench (full results).

Model	Model Size	short	medium	long	extra_long
LLMs (instruction-tuned)					
deepseek-llm-7b-chat	7B	0.185	0.131	0.096	0.072
llama3-8b-instruct	8B	0.109	0.082	**0.060**	0.055
llama3.1-8b-instruct	8B	**0.104**	**0.078**	0.061	**0.051**
qwen2.5-3b-instruct	3.1B	0.120	0.090	0.073	0.067
qwen2.5-7b-instruct	7.6B	0.110	0.082	0.069	0.051
qwen3-4b-instruct	4B	0.153	0.094	0.075	0.061
LLMs (base)					
deepseek-llm-7b-base	7B	0.161	0.105	0.084	0.084
llama3-8b	8B	**0.108**	**0.082**	0.064	**0.054**
llama3.1-8b	8B	0.110	0.082	**0.064**	0.054
qwen2.5-3b	3.1B	0.119	0.090	0.072	0.063
qwen3-4b	4B	0.128	0.086	0.070	0.058
Embedding models (instruction-tuned)					
e5-mistral-7b-instruct	7.1B	0.123	0.096	0.079	0.083
gte-qwen2-7b-instruct	7.6B	**0.109**	**0.078**	**0.066**	**0.054**
Embedding models					
all-minilm-l12-v2	33.4M	0.246	0.197	0.130	0.123
all-minilm-l6-v2	22.7M	0.248	0.186	0.121	0.119
all-mpnet-base-v2	109.5M	0.248	0.176	0.120	0.144
all-roberta-large-v1	355.4M	0.226	0.147	0.108	0.148
allenai-specter	110M	0.237	0.178	0.110	0.094
bert-base-uncased	110.1M	0.147	0.098	**0.075**	0.083
bge-base-en-v1.5	109.5M	0.309	0.222	0.130	0.151
bge-base-zh-v1.5	102M	0.200	0.143	0.096	0.089
bge-large-en-v1.5	335M	0.289	0.204	0.120	0.148
bge-large-zh-v1.5	326M	0.203	0.147	0.102	0.113
bge-m3	567M	0.354	0.296	0.180	0.155
bge-small-en-v1.5	33.4M	0.293	0.224	0.152	0.176
distiluse-base-multilingual-cased-v2	134.7M	0.275	0.255	0.158	0.092
e5-base-v2	109M	0.260	0.186	0.115	0.116
e5-large-v2	335.1M	0.245	0.190	0.131	0.137
e5-small-v2	33M	0.296	0.226	0.139	0.116
facebook-contriever	110M	0.259	0.168	0.134	0.206
facebook-contriever-msmarco	110M	0.275	0.202	0.144	0.184
gte-base	109.5M	0.274	0.182	0.123	0.158
gte-large	335.1M	0.251	0.170	0.109	0.131
gte-large-en-v1.5	409M	0.200	0.128	0.095	0.116
jina-embeddings-v2-base-en	137.4M	0.387	0.338	0.283	0.239
jina-embeddings-v2-small-en	33M	0.175	0.124	0.087	0.081
msmarco-distilbert-base-v4	66.4M	0.270	0.255	0.196	0.249
multilingual-e5-base	278M	0.197	0.171	0.120	0.083
multilingual-e5-large	559.9M	0.199	0.164	0.119	0.076
mxbai-embed-large-v1	335.1M	0.292	0.209	0.126	0.143
paraphrase-mpnet-base-v2	109.5M	0.260	0.189	0.124	0.170
paraphrase-multilingual-mpnet-base-v2	278M	0.360	0.297	0.208	0.202
qwen3-embedding-0.6b	595.8M	0.169	0.109	0.082	**0.048**
qwen3-embedding-4b	4B	0.178	0.111	0.085	0.058
qwen3-embedding-8b	7.6B	0.192	0.134	0.117	0.094
sfr-embedding-mistral	7.1B	**0.121**	**0.096**	0.079	0.083
snowflake-arctic-embed-l-v2	567.8M	0.273	0.235	0.143	0.072
Lexical / non-neural baselines					
ngram	–	**0.260**	**0.225**	**0.154**	**0.124**
ppm	–	0.331	0.308	0.254	0.274
tfidf	–	0.340	0.225	0.159	0.202

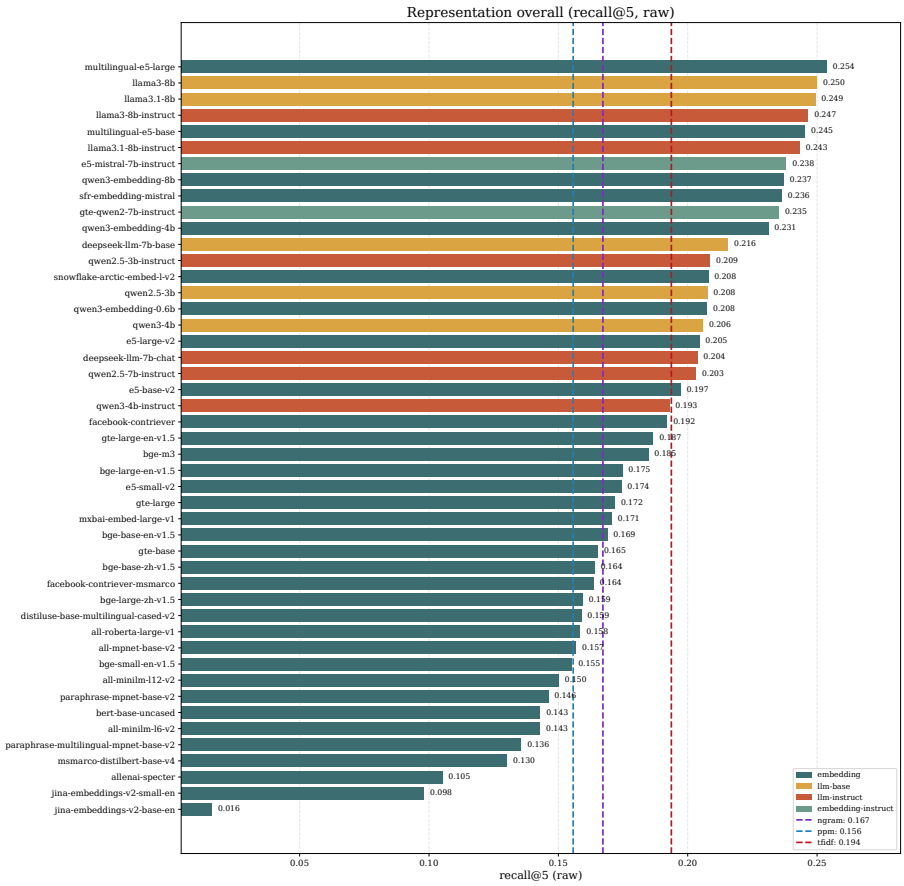

Figure 5: Overall AuthBench performance by Recall@5 (R@5).

Citation policy. When a dataset has an associated peer-reviewed publication, we treat it as the canonical reference (e.g., MARC for Amazon Reviews Multi; Babel Briefings for multilingual headlines). For mirrors or redistributed versions (e.g., Kaggle or Hugging Face hosting), we additionally cite the hosting page to ensure reproducibility via stable URLs and access dates.

No configured source is entirely absent from the current combined filtered output; however, the realized source distribution is highly skewed, with Exorde, Wikisource, Babel Briefings, and YouTube comments together contributing 74.4% of all documents.

E.1 Data Consent

AuthBench is derived from publicly accessible datasets and public-web sources released or exposed under documented provider licenses or site terms (Appendix E). We did not collect new data directly from individuals, and we did not contact data subjects. Where consent mechanisms are relevant, we rely on the original dataset providers' terms or the governing platform/site terms for public availability and redistribution. We further reduce privacy risk by anonymizing author identifiers (Section 3; Section 3.1) and applying conservative safety filtering (Appendix A.3).

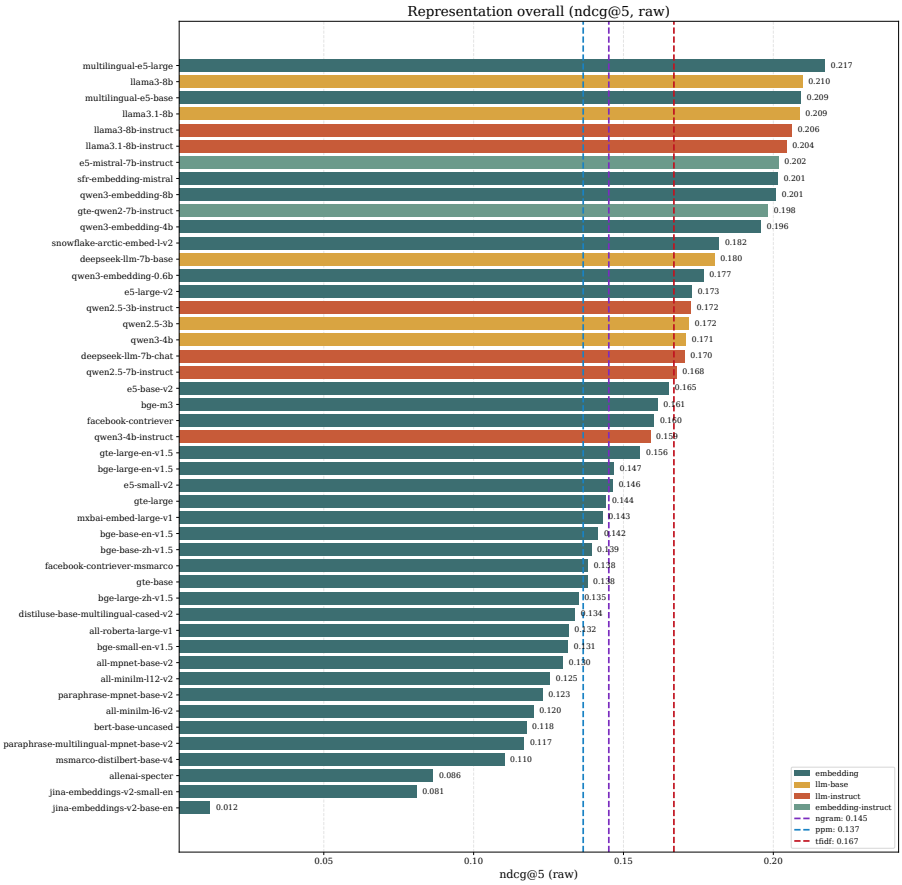

Figure 6: Overall AuthBench performance by nDCG@5.

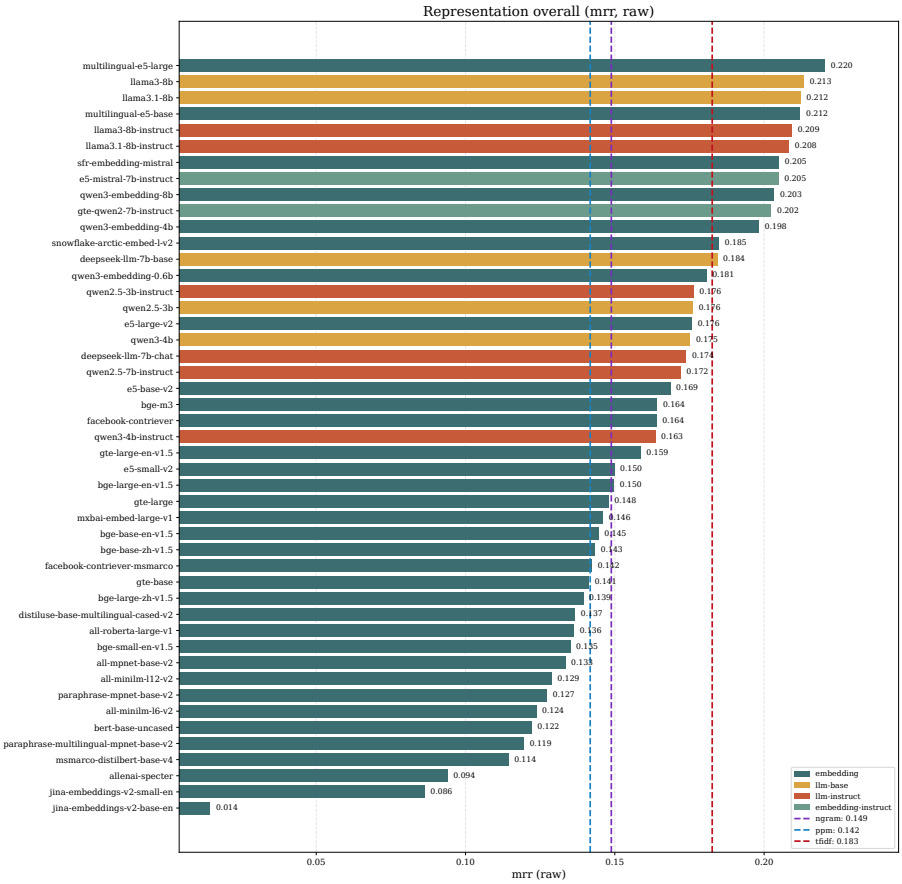

Figure 7: Overall AuthBench performance by MRR.

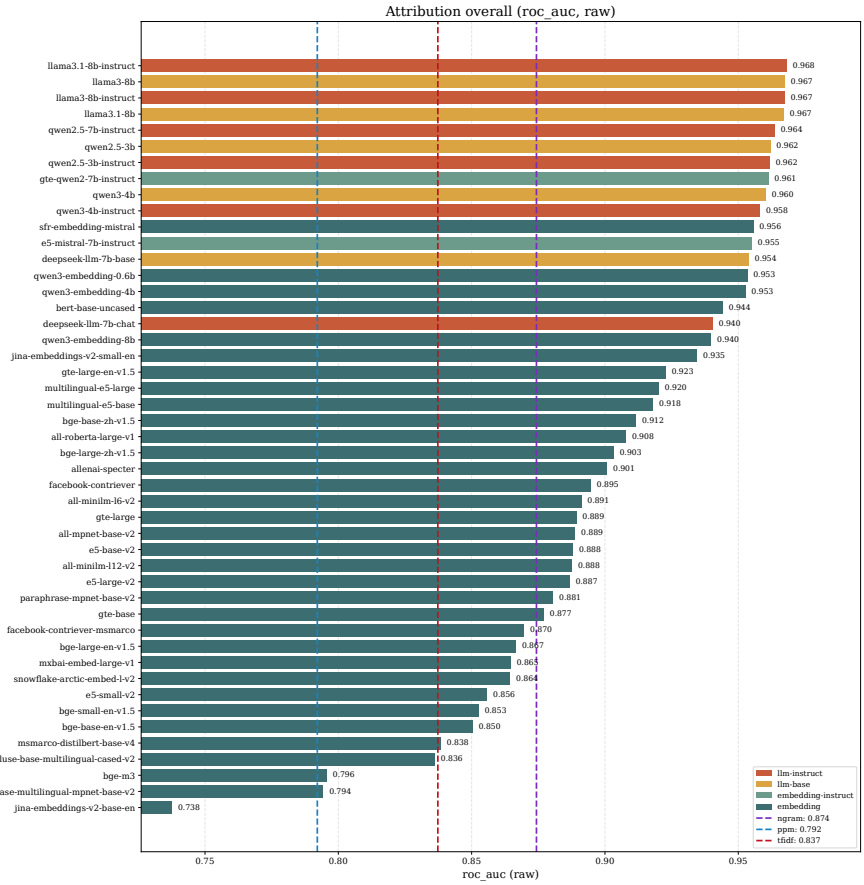

Figure 8: Overall AuthBench performance by ROC-AUC.

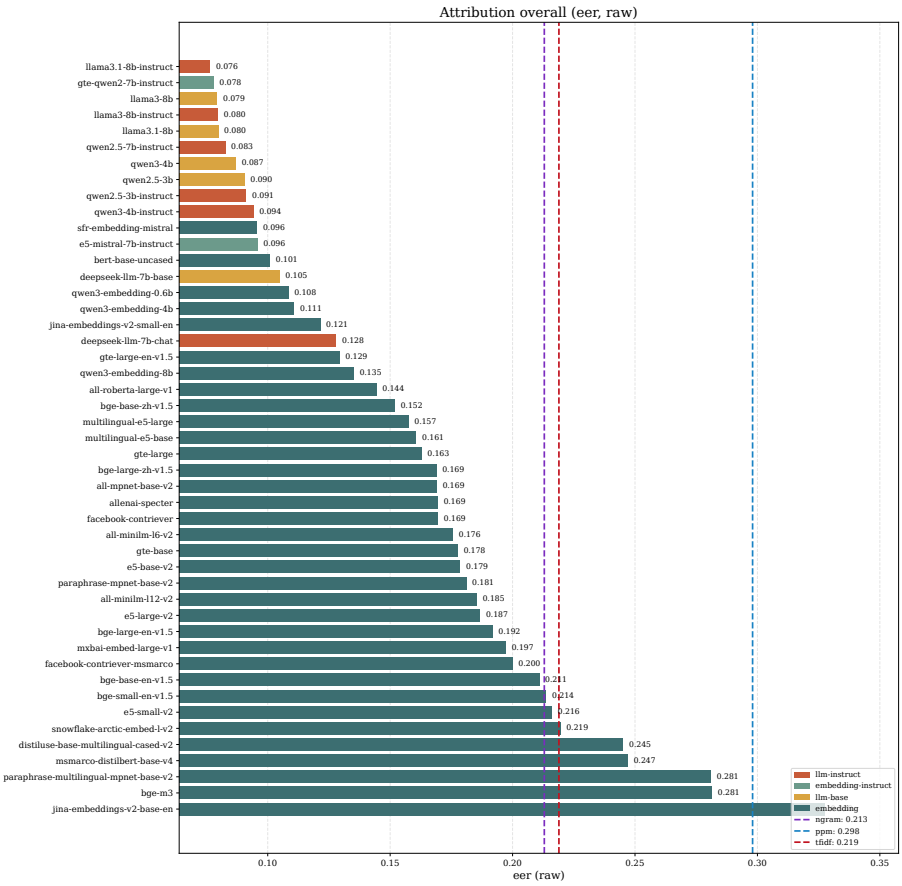

Figure 9: Overall AuthBench performance by EER.

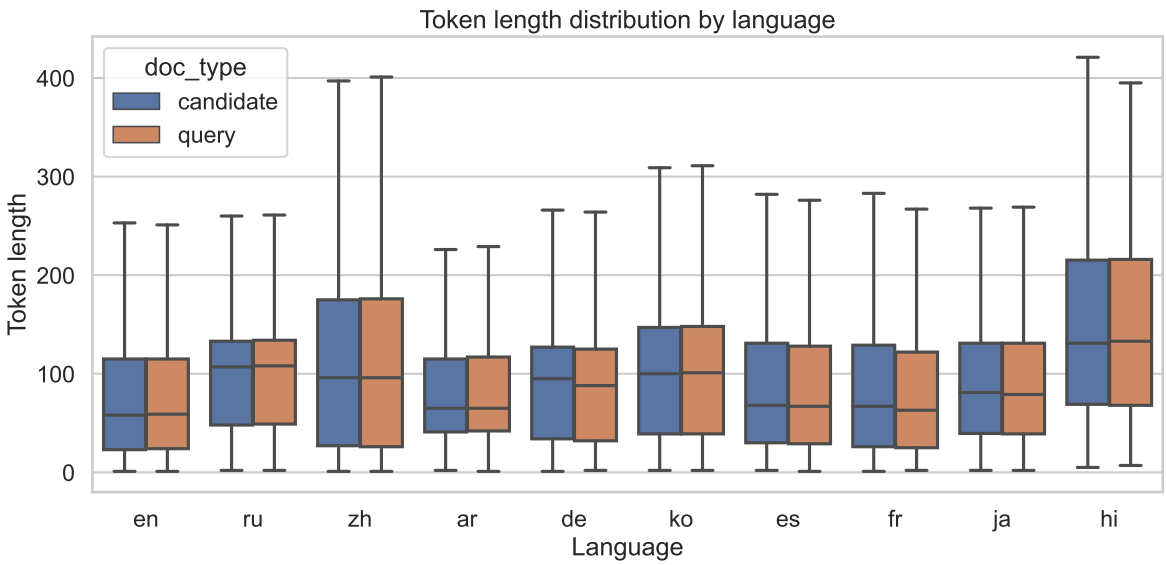

Figure 10: Token-length distribution per language. Each box plot summarizes the document-length distribution for one language in the current AuthBench release, highlighting cross-language differences in median length, spread, and long-tail behavior.

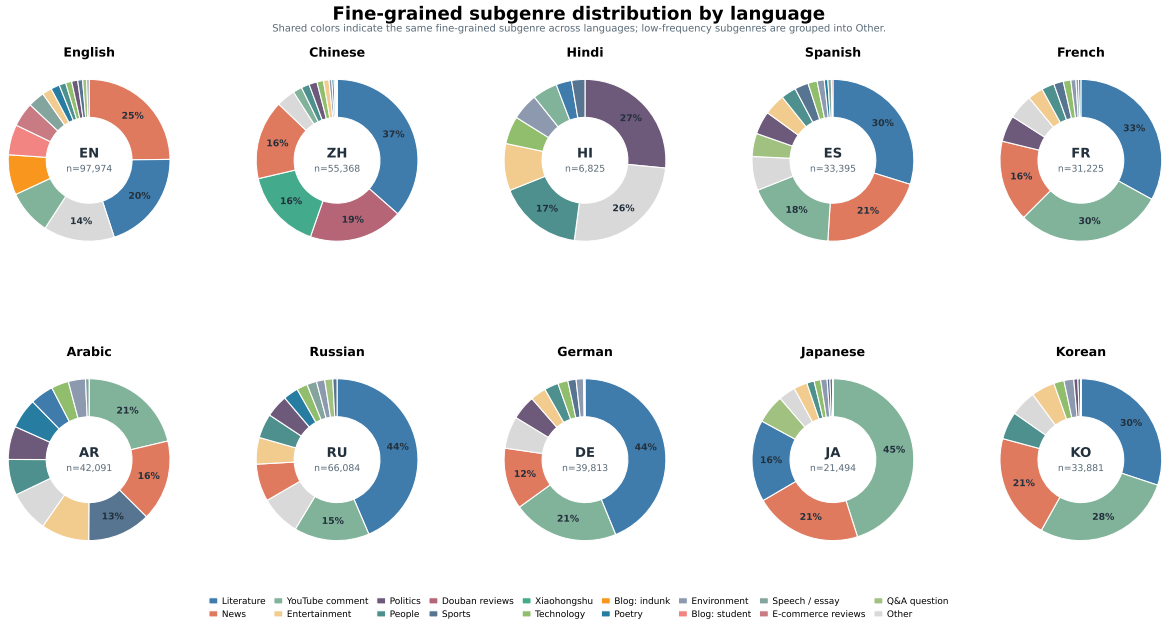

Figure 11: Fine-grained subgenre distribution by language. Each donut chart summarizes one language slice using a shared subgenre palette, while lower-frequency subgenres are grouped into `Other` to keep the figure legible.

Source	Lang.	Primary genre(s)	Scale	Author labels	Reference
Exorde	Multi	Social / news / forums	65M+/week	Yes (author hash)	(Exorde Labs, 2024)
Babel Briefings	30+	News headlines	4.7M	Partial (publisher / org)	(Leeb and Schölkopf, 2024; Leeb, 2023)
Amazon Reviews Multi (MARC)	6	E-commerce reviews	200k+/lang	Yes (reviewer IDs)	(Keung et al., 2020; mexwell, n.d.)
Blog Authorship Corpus	en	Blogs	681k posts	Yes	(Schler et al., 2006b; barilan, n.d.)
arXiv Abstracts (metadata snapshot)	en	Research papers	1.7M+	Yes (first author from metadata)	(arXiv, n.d.; Cornell University, n.d.)
Xiaohongshu / Weibo	zh	Social media	11k+	Yes (user IDs)	(yuanchunhong, n.d.)
Douban Reviews	zh	Media/book/music reviews	13.5M	Yes (reviewer IDs)	(fengzhujoey, n.d.)
Hindi Discourse Stories	hi	Literature (short stories)	53 stories	Yes	(Dhanwal et al., 2020; MIDAS Lab, IIIT-Delhi, n.d.)
Spanish PD Books	es	Literature	300k+ texts	Yes (metadata)	(PleIAs, 2024d)
French PD Books	fr	Literature	289k+ books	Yes (metadata)	(PleIAs, 2024a)
Arabic Classical Poetry	ar	Poetry	70k poems	Yes (poet)	(mdanok, n.d.)
Russian PD Corpus	ru	Literature/periodicals	8.5k titles	Yes (metadata)	(PleIAs, 2024c)
German PD Corpus	de	Literature/newspapers	260k+ texts	Yes (metadata)	(PleIAs, 2024b)
Stack Exchange Data Dump / API crawl	Multi	Q&A	Crawl-dependent; current processed corpus 6.5k docs	Yes (user IDs)	(Stack Exchange Meta, 2025)
Project Gutenberg	Multi	Literature / essays / speeches	Crawl-dependent; current processed corpus 23.6k docs	Yes (metadata author)	(Project Gutenberg)
Wikisource	Multi	Literature / drama / speeches	Crawl-dependent; current processed corpus 100.8k docs	Yes (author metadata / heuristics)	(Wikimedia Dumps)
YouTube comments	Multi	Social media comments	Crawl-dependent; current processed corpus 104.7k comments	Yes (comment author)	(Google Developers)

Table 15: Raw sources used for the current AuthBench release. "Scale" reflects either the approximate size of the upstream source or, for crawl-dependent sources, the size of the processed corpus produced by the current collection setting before final benchmark selection.

Source	#Docs	Share
Exorde	94,231	22.0%
Wikisource	78,984	18.4%
Babel Briefings	73,676	17.2%
YouTube comments	71,808	16.8%
Blog Authorship Corpus	22,494	5.3%
Project Gutenberg	18,739	4.4%
Russian PD Corpus	12,728	3.0%
Douban Reviews	10,424	2.4%
Xiaohongshu / Weibo	8,869	2.1%
French PD Books	8,761	2.0%
German PD Corpus	8,400	2.0%
Spanish PD Books	4,961	1.2%
Amazon Reviews Multi (MARC)	4,924	1.2%
Stack Exchange Data Dump / API crawl	4,651	1.1%
Arabic Classical Poetry	2,503	0.6%
arXiv Abstracts (metadata snapshot)	1,784	0.4%
Hindi Discourse Stories	213	0.0%

Table 16: Source composition of the current combined AuthBench export. All 17 configured sources are represented in the materialized benchmark; the table reports their realized document counts after filtering, redundancy reduction, cross-phase merge cleanup, and final selection.

Table 17: Licensing / terms and release mode by source. "Release mode" indicates whether we redistribute normalized text (Tier A) or provide manifest-only reconstruction (Tier B). This table covers all 17 sources in the current release.

Source	**License / terms (summary)**	**Release mode**	**Notes / pointer**
Exorde	MIT (dataset card)	Tier A	Provider page explicitly states MIT.
Babel Briefings	CC BY-NC-SA 4.0 (dataset card)	Tier A	Attribution, non-commercial use, and share-alike obligations apply.
Amazon Reviews Multi (MARC)	Provider-restricted / non-permissive redistribution; dataset card does not expose a standard open license	Tier B	Safer to require reconstruction from the original release rather than redistribute normalized text.
Blog Authorship Corpus	No clear standardized redistribution license found in the public host materials	Tier B	Conservative manifest-only treatment is safer than direct text redistribution.
arXiv Abstracts (metadata snapshot)	Metadata access is available via arXiv services, but article licenses vary by paper	Tier B	We treat this as metadata-derived text and do not redistribute PDFs or assume one uniform article license.
Xiaohongshu / Weibo	Platform / provider terms for redistribution are unclear	Tier B	Manifest-only unless explicit redistribution permission is documented.
Douban Reviews	Platform / provider terms for redistribution are unclear	Tier B	Same conservative handling as other platform-hosted review data.
Hindi Discourse Stories	Redistribution terms are unclear from public materials	Tier B	Treat as a research-use source unless explicit permission is documented.
Spanish PD Books	Public-domain source texts	Tier A	Normalized text can be redistributed as public-domain material.
French PD Books	Public-domain source texts	Tier A	Normalized text can be redistributed as public-domain material.
Arabic Classical Poetry	Public host available, but redistribution terms should be verified per host copy	Tier B	Conservative default is manifest-only unless an explicit redistributable license is documented.
Russian PD Corpus	Public-domain source texts	Tier A	Normalized text can be redistributed as public-domain material.
German PD Corpus	Public-domain source texts	Tier A	Normalized text can be redistributed as public-domain material.
Stack Exchange Data Dump / API crawl	CC BY-SA for public user contributions (version depends on post date; newer content is CC BY-SA 4.0)	Tier A	Attribution and share-alike requirements apply to redistributed text.
Project Gutenberg	Mostly U.S.-public-domain texts, but Project Gutenberg license / trademark terms still apply	Tier B	Conservative manifest-only handling avoids over-claiming unrestricted redistribution across jurisdictions and formats.
Wikisource	Usually CC BY-SA with possible page-level public-domain variation depending on the work	Tier B	Attribution / share-alike obligations and page-level rights should be tracked before direct redistribution.
YouTube comments	Governed by YouTube API Terms and Developer Policies rather than a simple open-content license	Tier B	Use manifests / derived metadata only; avoid direct text redistribution by default.