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

Anonymous ACL submission

Abstract

We introduce AuthBench, a large-scale multilingual benchmark for Authorship Representation, enabling evaluation of authorship Attribution and Verification across numerous genres and lengths. AuthBench contains 293,029 documents of diverse lengths (short, medium, long, and extra-long), authored by 53,281 individuals across ten widely used languages (en, zh, hi, es, fr, ar, ru, de, ja, ko), spanning 9 primary genres with 62 fine-grained subgenres. AuthBench supports two complementary evaluations: (i) *authorship attribution*, operationalized as same-author document retrieval, scored by Success@K, Recall@K, and nDCG@K; and (ii) *authorship verification*, i.e., same-author binary decisions over query–candidate pairs, scored by equal error rate (EER). We comprehensively evaluate state-of-the-art (SOTA) instruction-tuned and embedding models on AuthBench. Experiments with SOTA models show that performance remains far from saturated: the best Success@5 reaches only 0.542, and the best overall EER is 0.078. Moreover, SOTA models exhibit substantial performance variation across languages, document lengths, and genres, highlighting persistent challenges for robust authorship modeling. We release AuthBench along with its evaluation toolkit anonymously at `https://anonymous.4open.science/r/AURA_Bench-366E/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 benchmarks fall short of comprehensively evaluating these capabilities in authorship representation, as they are typically restricted to single languages or genres. For instance, widely used datasets like the Blog Authorship Corpus (Schler et al., 2006a), Enron emails (Klimt and Yang, 2004), and PAN[1] shared tasks (Potthast et al., 2016; Bevendorff et al., 2020) are limited to specific domains. Similarly, other benchmarks focusing on scientific writing (Bevendorff et al., 2023), cross-domain settings (e.g., LUAR) (Rivera-Soto et al., 2021), or journalist-centric pairings (Ma et al., 2025) remain narrowly scoped. Furthermore, data distributions in these benchmarks are often highly skewed: for example, scientific authorship datasets predominantly feature documents exceeding 500 words. Finally, current benchmarks lack a unified framework to support both *retrieval-style authorship representation* to find documents by the same author in a collection of documents and *verification-style authorship attribution* to confirm that a given pair of documents is by the same author.

To overcome the above limitations of existing benchmarks on authorship representation, we introduce AuthBench, a large-scale multilingual benchmark for Authorship Representation enabling Attribution and Verification across genres

[1] Each edition of PAN focuses on one specific genre.

Benchmark	Multilingual	Multi-genre	Length control	Large-scale	Standard splits	Attribution	Verification
PAN shared tasks	✓	✓	✗	✗	✓	✓	✓
Blog Authorship Corpus	✗	✗	✗	✓	✗	✗	✗
Enron Emails	✗	✗	✗	✓	✗	✗	✗
MARC	✓	✗	✗	✓	✓	✗	✗
SMAuC	✗	✗	✗	✓	✗	✗	✗
CROSSNEWS	✗	✓	✗	✓	✓	✗	✓
AIDBench	✗	✓	✗	✗	✓	✓	✓
AuthBench (ours)	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of our AuthBench with existing authorship benchmarks.

and document lengths. AuthBench is constructed from 13 public data sources (see Appendix E for details). We normalize all data into a unified format, filter out low-quality samples, and perform stratified sampling to balance genre and length distributions across languages. Finally, we conduct additional automated checks and manual inspections to ensure the overall quality of AuthBench. As a result, AuthBench comprises 293,029 documents of diverse lengths (short, medium, long, and extra-long), authored by 53,281 individuals across ten widely used languages (en, zh, hi, es, fr, ar, ru, de, ja, ko), spanning nine primary genres with 62 fine-grained subgenres. AuthBench supports two complementary evaluation tasks: (i) *authorship attribution*, formulated as a retrieval task evaluated with ranking-based metrics (Success@K, Recall@K, nDCG@K); and (ii) *authorship verification*, formulated as binary decision over query–candidate pairs and evaluated using threshold-based metrics (EER). We release standardized dataset splits and a unified evaluation toolkit to facilitate reproducible and comprehensive evaluation.

We conduct a comprehensive evaluation of the SOTA models on AuthBench, including embedding, instructed, and base LLMs. The evaluation results show that the best overall model performance of Success@5 on authorship attribution is 0.542, and the best overall model performance of EER on authorship attribution is 0.078. Moreover, SOTA models exhibit substantial performance variation 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 (often framed as closed-set classification or large-scale retrieval), and (ii) authorship verification, which determines whether two documents were written by the same author. 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 works, including LUAR (Rivera-Soto et al., 2021), style-aware embedding methods (Wang and Iyyer, 2023), and stylistic disentanglement analyses (Michailidis et al., 2022), operationalize attribution via embedding similarity and retrieval-style evaluation (e.g., Success@K, Recall@K, nDCG@K), and evaluate verification using pairwise decision metrics such as equal error rate (EER).

Authorship Benchmark Development. Early authorship benchmarks were largely built on small-scale, monolingual, and domain-specific corpora, as summarized in foundational surveys (Stamatatos, 2009; Neal et al., 2017). Datasets such as the Blog Authorship Corpus (Schler et al., 2006a), the Enron email dataset (Klimt and Yang, 2004), and later English-centric collections used in neural modeling studies (Ruder et al., 2016; Jafariakinabad and Hua, 2019) supported important methodological progress, but provided limited coverage in terms of language diversity, genre breadth, and document length control. Besides, none of them provides the standard evaluation protocols, making it hard to evaluate current SOTA models. The PAN shared tasks subsequently established standardized evaluation protocols for authorship-related problems (Bevendorff et al., 2020). While influential, individual PAN editions typically addressed a single task formulation and relied on datasets restricted to specific languages and genres. More recent work has sought to broaden authorship evaluation beyond these settings by in-

troducing more domain-oriented authorship corpora, such as scientific writing datasets (SMAuC) (Bevendorff et al., 2023) and journalist news–tweet pairings (CROSSNEWS) (Ma et al., 2025). In parallel, LLM-focused authorship benchmarks (e.g., AIDBench) (Wen et al., 2024) and studies examining authorship risks in large language models (Huang et al., 2024a,b) have emerged to evaluate the flaws of the SOTA models on authorship. Nevertheless, despite this growing body of work, existing benchmarks remain predominantly domain-specific, often exhibit highly imbalanced author distributions, and lack unified control over language, genre, and document length. We provide a comparison of our Authbench with existing benchmarks in Table 1.

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 1 illustrates the construction workflow of AuthBench. We constructed the benchmark using 13 publicly available data sources. Descriptions of these datasets are available in Appendix E. To ensure consistency across heterogeneous corpora, we first normalize each source into a unified record format containing the following fields: language, genre, source ID, anonymized author ID, document content, and token length. We provide examples in Figure 2. We then remove redundancy using a combination of exact hashing and near-duplicate detection. To further improve data quality, we apply safety and quality filters to exclude low-quality or problematic documents. Next, to achieve balanced distributions, we perform bucket-based sampling across languages, document length ranges, and genres, while enforcing per-author caps to prevent dominance by highly prolific writers. Finally, to produce leakage-resistant data splits, we apply stratified partitioning with deduplication-aware assignment, ensuring that duplicate or near-duplicate documents do not appear across different splits. We additionally conduct automated checks and manual inspections to validate the quality of the final AuthBench. Automated checks verify schema integrity, bucket assignment, and split hygiene (no author overlap and no duplicate or near-duplicate leakage across splits). For manual inspection, the authors randomly sample documents across languages, genres, and length buckets to screen for noise, boilerplate, corrupted text, and unsafe content. Complete construction details are provided in Appendix A.

Category	#Docs	Share
Overall		
Total	293,029	100.0%
Languages	10	–
Authors	53,281	–
Avg. docs per author	2.46	–
Genres (fine-grained)	62	–
By language		
English (en)	104,786	35.8%
Spanish (es)	29,891	10.2%
Chinese (zh)	28,487	9.7%
French (fr)	23,865	8.1%
German (de)	23,864	8.1%
Arabic (ar)	23,602	8.1%
Russian (ru)	21,703	7.4%
Japanese (ja)	14,331	4.9%
Korean (ko)	14,178	4.8%
Hindi (hi)	8,322	2.8%
By length		
Short	32,143	11.0%
Medium	178,191	60.8%
Long	79,125	27.0%
Extra-long	3,570	1.2%

Table 2: Statistics of our AuthBench.

3.2 Statistics of AuthBench

After construction, AuthBench comprises 293,029 documents authored by 53,281 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`, `news`, `blog`, `literature`, `media_reviews`, `ecommerce_reviews`, `poetry`, `research_paper`. 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. Additional detailed statistics are provided in Appendix D.

Figure 1: AuthBench construction pipeline. Full details are provided in Appendix A.


```
"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":
"由于中国政府加大对游戏内容的管控力度，
苹果公司正将数千款游戏应用从其中国平台」
上下架...",
"source": "babel_briefings",
"token_length": 74
```


Figure 2: Examples in our AuthBench.

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; alternatives such as CLS/last are supported where appropriate), and L2-normalize the resulting embedding. Similarities are computed with the cosine distance between normalized embeddings. We use batch size 32 and the full-sized within-split candidate pool; verification EER uses sampled 50 negatives per query as defined in Appendix A.8.

Models. We evaluate a range of SOTA base, embedding, and instruction-tuned models. Full model details are provided in Table 7 of Appendix A.8.

Metrics. AuthBench supports both authorship attribution and authorship verification tasks. For authorship attribution, we report Success@5 (S@5), Recall@5 (R@5), and nDCG@5. For authorship verification, we report EER. Formal definitions of all evaluation metrics are provided in Appendix A.8.

5 Experimental Results

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 demonstrate the training dynamics and the effectiveness of fine-tuning on AuthBench to enhance authorship representation.

5.1 Main Results

Finding 1: base LLMs consistently provide the strongest authorship signals, outperforming instruction-tuned and embedding-based models. Table 3 summarizes the overall performance of a subset of SOTA models on AuthBench. Due to space constraints, we report results for all evaluated models in Appendix B. We categorize the evaluated models into four groups: instruction-tuned LLMs, base LLMs, instruction-tuned embedding models, and base embedding models.

Overall, base LLMs consistently outperform all other model groups, often surpassing their instruction-tuned counterparts. For example, `llama3-8b` achieves the best overall authorship attribution performance, with scores of 0.542 (S@5), 0.537 (R@5), and 0.441 (nDCG@5), outperforming `llama3-8b-instruct` across all metrics. However, model scale is not the sole determinant of performance. Notably, the compact `qwen2.5-3b` outperforms the significantly larger `deepseek-llm-7b-base` (S@5: 0.496 vs. 0.479), highlighting the efficiency of recent architectures in capturing stylistic features despite limited parameter counts.

While base LLMs dominate attribution metrics, instruction-tuned LLMs remain competitive in the verification task. Remarkably,

Model	Size	S@5 ↑	R@5 ↑	nDCG@5 ↑	EER ↓
Instruction-tuned LLMs					
deepseek-coder-6.7b-instruct	6.7B	0.455	0.450	0.368	0.095
deepseek-llm-7b-chat	7B	0.446	0.439	0.354	0.107
llama3-8b-instruct	8B	**0.530**	**0.524**	**0.428**	0.086
llama3.1-8b-instruct	8B	0.527	0.520	0.427	0.085
qwen2.5-3b-instruct	3.1B	0.498	0.491	0.396	**0.078**
qwen2.5-7b-instruct	7.6B	0.487	0.481	0.390	0.081
qwen3-4b-instruct	4B	0.483	0.476	0.381	0.089
Base LLMs					
deepseek-llm-7b-base	7B	0.479	0.473	0.379	0.095
llama3-8b	8B	**0.542**	**0.537**	**0.441**	0.089
llama3.1-8b	8B	0.534	0.528	0.436	0.089
qwen2.5-3b	3.1B	0.496	0.488	0.394	**0.081**
qwen3-4b	4B	0.481	0.475	0.385	0.089
Instruction-tuned embedding models					
e5-mistral-7b-instruct	7.1B	0.467	0.460	0.384	0.099
gte-qwen2-7b-instruct	7.6B	**0.490**	**0.484**	**0.398**	**0.089**
sfr-embedding-mistral	7.1B	0.468	0.462	0.383	0.099
Base embedding models					
all-minilm-l12-v2	33M	0.322	0.318	0.254	0.161
all-minilm-l6-v2	23M	0.322	0.318	0.251	0.149
all-mpnet-base-v2	109M	0.346	0.340	0.272	0.144
all-roberta-large-v1	355M	0.380	0.374	0.296	0.127
allenai-specter	110M	0.312	0.310	0.251	0.152
bert-base-uncased	110M	**0.394**	**0.388**	**0.311**	**0.103**
bge-base-en-v1.5	109M	0.360	0.355	0.282	0.172
Lexical baseline					
tfidf	n/a	**0.326**	**0.321**	**0.264**	**0.177**

Table 3: Main Results on AuthBench (subset). **Bold** = best, underlined = second best within each model group.

qwen2.5-3b-instruct achieves the lowest EER of 0.078 across all evaluated models. This suggests that while instruction tuning might degrade attribution performance (e.g., S@5), it may preserve or even enhance the model's decision boundaries for binary authorship verification (i.e., EER).

In contrast, instruction-tuned embedding models exhibit noticeably weaker attribution performance than base LLMs. Finally, base embedding models perform the worst overall due to the small model size. Strikingly, traditional dense retrievers often struggle to capture stylistic nuances beyond lexical overlap. As shown in the table, several base embedding models (e.g., allenai-specter) perform comparably to—or even worse than—the simple TF-IDF baseline (S@5: 0.326), further underscoring the necessity of LLM-based models for authorship representation.

5.2 Results by Language

Finding 2: model performance varies substantially across languages. Table 4 reports S@5 results for a subset of representative state-of-the-art models across languages on AuthBench. Due to space constraints, we report results for all evaluated models and additional metrics Appendix B. Overall, model performance exhibits pronounced language-dependent variation. For example, llama3-8b achieves an S@5 of approximately 0.824 on Japanese but only 0.335 on Arabic. Similarly, all-roberta-large-v1 reaches 0.575 on Hindi while dropping to 0.283 on Korean, yielding performance gaps of up to 0.5 across languages. We further observe that larger models (e.g., those with more than 3B parameters) appear to saturate performance on Japanese, whereas smaller models perform substantially worse (often below 0.5). In contrast, for most other languages, even the strongest SOTA models remain far from saturation, indicating significant room for future improvement in multilingual authorship modeling.

Model	Size	ar	de	en	es	fr	hi	ja	ko	ru	zh
LLMs (instruction-tuned)											
deepseek-coder-6.7b-instruct	6.7B	0.305	0.662	0.410	0.535	0.581	0.525	**0.941**	0.435	0.391	0.504
deepseek-llm-7b-chat	7B	0.266	0.620	0.500	0.564	0.516	0.575	**0.941**	0.402	0.323	0.488
llama3-8b-instruct	8B	0.305	**0.761**	**0.566**	0.644	0.629	0.800	0.765	0.500	**0.504**	0.504
llama3.1-8b-instruct	8B	0.310	0.746	0.530	**0.693**	**0.645**	**0.850**	0.647	**0.511**	0.474	0.504
qwen2.5-3b-instruct	3.1B	**0.350**	0.648	0.464	0.594	0.629	0.600	0.765	0.478	0.459	**0.543**
qwen2.5-7b-instruct	7.6B	0.335	0.648	0.458	0.574	0.597	0.650	0.882	0.424	0.444	**0.543**
qwen3-4b-instruct	4B	0.340	0.662	0.434	0.564	0.581	0.700	0.765	0.457	0.421	**0.543**
LLMs (base)											
deepseek-llm-7b-base	7B	0.266	0.648	0.524	0.634	0.613	0.675	**0.941**	0.424	0.361	0.520
llama3-8b	8B	0.335	**0.775**	**0.560**	0.663	**0.645**	**0.850**	0.824	**0.500**	**0.496**	0.512
llama3.1-8b	8B	0.335	**0.775**	0.554	**0.673**	**0.645**	0.800	0.765	0.478	0.489	0.496
qwen2.5-3b	3.1B	**0.345**	0.634	0.458	0.594	0.629	0.600	0.882	0.489	0.444	**0.543**
qwen3-4b	4B	0.340	0.676	0.464	0.574	0.565	0.625	0.706	0.446	0.406	0.535
Embedding models (instruction-tuned)											
e5-mistral-7b-instruct	7.1B	0.276	**0.648**	0.476	**0.604**	0.532	0.625	**0.882**	**0.457**	0.451	0.441
gte-qwen2-7b-instruct	7.6B	**0.296**	0.634	**0.488**	0.564	**0.645**	**0.675**	0.765	**0.457**	**0.474**	**0.535**
sfr-embedding-mistral	7.1B	0.271	0.634	**0.488**	**0.604**	0.548	0.650	0.706	**0.457**	0.444	0.465
Embedding models											
all-minilm-l12-v2	33M	0.163	0.380	0.386	0.396	0.387	0.600	0.471	0.326	0.248	0.339
all-minilm-l6-v2	23M	0.192	0.408	0.380	0.436	0.306	0.475	0.294	0.293	0.256	0.370
all-mpnet-base-v2	109M	0.192	0.507	**0.470**	0.406	0.323	0.525	0.353	0.283	0.248	0.394
all-roberta-large-v1	355M	**0.266**	**0.634**	0.428	0.465	**0.419**	0.575	**0.647**	0.283	0.263	0.370
allenai-specter	110M	0.167	0.577	0.343	0.455	0.403	0.325	0.294	0.326	0.241	0.260
bert-base-uncased	110M	0.212	0.563	0.458	**0.495**	0.403	0.575	0.588	**0.424**	**0.316**	**0.402**
bge-base-en-v1.5	109M	0.197	0.592	0.386	0.455	0.371	**0.625**	0.588	0.348	0.286	0.346
Lexical baseline											
tfidf	n/a	**0.291**	**0.310**	**0.313**	**0.267**	**0.419**	**0.575**	**0.529**	**0.380**	**0.293**	**0.299**

Table 4: Results by language (S@5).

Model	Size	blog	ecommerce_reviews	literature	media_reviews	news	poetry	research_paper	social_media
LLMs (instruction-tuned)									
deepseek-coder-6.7b-instruct	6.7B	0.146	0.308	0.575	0.218	0.448	0.288	0.840	0.503
deepseek-llm-7b-chat	7B	0.174	0.333	0.511	0.164	0.462	0.308	**0.880**	0.514
llama3-8b-instruct	8B	**0.416**	0.385	**0.632**	0.200	**0.560**	0.288	**0.880**	0.592
llama3.1-8b-instruct	8B	0.412	0.359	**0.632**	0.200	0.538	0.308	**0.880**	**0.594**
qwen2.5-3b-instruct	3.1B	0.218	0.333	0.592	0.236	0.538	**0.327**	0.800	0.552
qwen2.5-7b-instruct	7.6B	0.156	**0.410**	0.586	0.236	0.513	0.288	0.760	0.585
qwen3-4b-instruct	4B	0.145	0.359	0.575	**0.255**	0.473	0.308	0.800	0.568
LLMs (base)									
deepseek-llm-7b-base	7B	0.323	0.385	0.569	0.200	0.487	0.250	**0.840**	0.540
llama3-8b	8B	**0.412**	0.385	0.626	0.182	0.574	**0.308**	**0.840**	**0.623**
llama3.1-8b	8B	0.351	0.333	**0.632**	0.182	**0.581**	**0.308**	**0.840**	0.600
qwen2.5-3b	3.1B	0.218	0.333	0.586	**0.255**	0.531	**0.308**	0.800	0.552
qwen3-4b	4B	0.273	**0.410**	0.575	**0.255**	0.502	**0.308**	0.800	0.530
Embedding models (instruction-tuned)									
e5-mistral-7b-instruct	7.1B	0.240	0.333	0.615	0.127	0.520	0.173	**0.840**	0.509
gte-qwen2-7b-instruct	7.6B	**0.251**	0.333	0.615	**0.236**	**0.538**	**0.250**	0.800	**0.542**
sfr-embedding-mistral	7.1B	0.246	**0.359**	**0.621**	0.164	0.527	0.154	**0.840**	0.503
Embedding models									
all-minilm-l12-v2	33M	0.200	0.128	0.391	0.109	0.256	0.058	**0.880**	0.407
all-minilm-l6-v2	23M	0.195	0.154	0.408	0.127	0.285	0.038	0.840	0.401
all-mpnet-base-v2	109M	**0.407**	0.256	0.414	**0.164**	0.296	0.096	0.840	0.379
all-roberta-large-v1	355M	0.207	**0.333**	0.437	**0.164**	0.310	**0.269**	0.840	0.416
allenai-specter	110M	0.218	0.154	0.466	0.055	0.292	0.058	0.800	0.351
bert-base-uncased	110M	0.241	0.282	**0.540**	0.127	**0.397**	0.115	0.800	0.409
bge-base-en-v1.5	109M	0.223	0.256	0.489	0.145	0.285	0.135	**0.880**	**0.444**
Lexical baseline									
tfidf	n/a	**0.172**	**0.103**	**0.379**	**0.073**	**0.310**	**0.250**	**0.800**	**0.385**

Table 5: Results by genre (S@5).

5.3 Results by Genre

Finding 3: model performance varies significantly by domain, yet the superiority of LLM-based approaches remains consistent across nearly all genres. Table 5 presents a fine-grained breakdown of model performance across eight major genres. Due to space constraints, we report results for all evaluated models and additional metrics in Appendix B. Authorship signals appear strongest in structured or formal do-

Model	Size	short	medium	long	extra_long
LLMs (instruction-tuned)					
deepseek-coder-6.7b-instruct	6.7B	0.333	0.571	0.402	0.267
deepseek-llm-7b-chat	7B	0.370	0.532	0.403	0.320
llama3-8b-instruct	8B	0.444	0.607	**0.491**	**0.427**
llama3.1-8b-instruct	8B	**0.481**	**0.610**	0.489	0.373
qwen2.5-3b-instruct	3.1B	0.370	0.589	0.461	0.333
qwen2.5-7b-instruct	7.6B	0.444	0.563	0.459	0.307
qwen3-4b-instruct	4B	0.370	0.592	0.434	0.307
LLMs (base)					
deepseek-llm-7b-base	7B	0.370	0.571	0.442	0.307
llama3-8b	8B	**0.407**	0.623	**0.511**	**0.387**
llama3.1-8b	8B	**0.407**	**0.625**	0.493	**0.387**
qwen2.5-3b	3.1B	**0.407**	0.579	0.463	0.333
qwen3-4b	4B	**0.407**	0.584	0.432	0.320
Embedding models (instruction-tuned)					
e5-mistral-7b-instruct	7.1B	**0.370**	0.571	0.417	0.320
gte-qwen2-7b-instruct	7.6B	0.333	**0.581**	**0.455**	0.320
sfr-embedding-mistral	7.1B	**0.370**	0.579	0.411	**0.333**
Embedding models					
all-minilm-l12-v2	33M	0.370	0.344	0.300	0.347
all-minilm-l6-v2	23M	0.333	0.354	0.296	0.333
all-mpnet-base-v2	109M	**0.444**	0.351	0.327	**0.413**
all-roberta-large-v1	355M	**0.444**	0.398	**0.371**	0.333
allenai-specter	110M	0.296	0.398	0.256	0.267
bert-base-uncased	110M	0.333	**0.494**	0.333	0.333
bge-base-en-v1.5	109M	0.407	0.437	0.314	0.267
Lexical baseline					
tfidf	n/a	**0.444**	**0.424**	**0.262**	**0.227**

Table 6: Results by Length (S@5).

mains. In the research_paper category, almost all evaluated models, including the lexical TF-IDF baseline, achieve a pretty good score above 0.8, likely due to highly distinctive individual vocabularies in academic writing. Conversely, media_reviews and ecommerce_reviews emerge as the most challenging domains, with the best-performing qwen3-4b-instruct only reaching 0.255 on the former. This suggests that short, opinionated texts with high cross-author lexical overlap pose a significant challenge for stylistic differentiation. Moveover, base LLMs, particularly the llama3 family, demonstrate a decisive advantage in more ”creative” genres such as literature, poetry, and blog. For instance, in literature, llama3.1-8b achieves an S@5 of 0.632, substantially outperforming the best embedding model (bge-base-en-v1.5 at 0.489) and the TF-IDF baseline (0.379). This indicates that the deep contextual representations in LLMs are better at capturing the complex syntactic and narrative nuances that define literary style compared to shallow embedding or lexical features.

5.4 Performance by Length

Within each model scale, large models perform better on medium **length documents, whereas small models perform better on** short **documents. Moreover, when comparing across scales, small models achieve comparable or even superior performance to large models on** short **documents.** Table 6 presents a fine-grained breakdown of model performance across four document lengths. Due to space constraints, we report results for all evaluated models and additional metrics in Appendix B. Performance for nearly all LLM-based models peaks in the medium length category. For instance, llama3.1-8b jumps from 0.407 on short documents to 0.625 on medium documents, before declining on longer inputs. This suggests that medium-length documents provide an optimal balance: they contain sufficient stylistic markers (unlike short texts) without introducing the noise or context dilu-

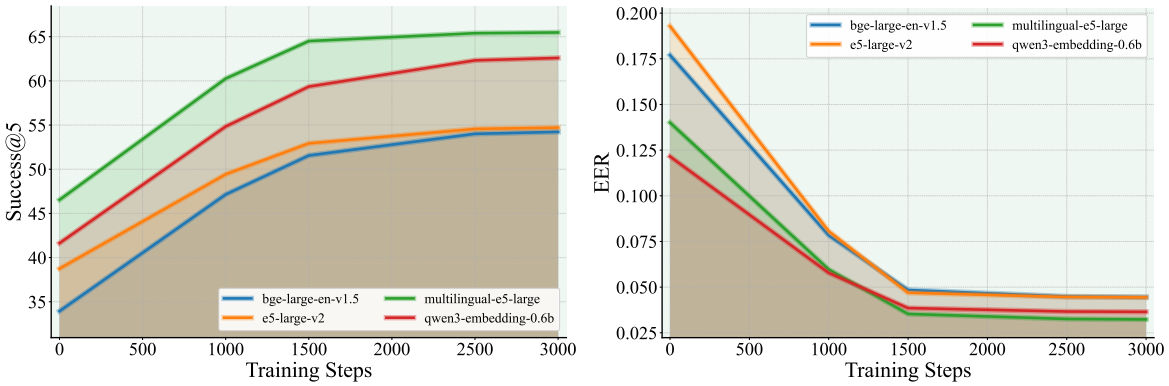

Figure 3: Training dynamics of select models.

tion often found in `long` or `extra_long` documents. Interestingly, small models achieve competitive performance on `short` documents compared to significantly larger LLMs. For example, `all-mpnet-base-v2` achieves a score of 0.444 with only 109M parameters, rivaling the 8B-parameter `llama3.1-8b-instruct` (0.481). This suggests that in short document regimes, authorship attribution relies less on complex semantic modeling and more on local stylistic features or lexical patterns, which compact embedding models can extract highly efficiently.

Base LLMs demonstrate superior stability as text length increases compared to embedding models. On `extra_long` texts, `llama3-8b-instruct` achieves the highest score of 0.427, maintaining strong performance where others falter. In sharp contrast, the lexical `TF-IDF` baseline collapses on longer texts, dropping from 0.444 (`short`) to a dismal 0.227 (`extra_long`). This confirms that while simple lexical statistics suffice for short documents, deep contextual understanding is a prerequisite for modeling the coherent stylistic patterns of lengthy documents.

5.5 Training Dynamics

To demonstrate the utility of AuthBench as a training resource, we fine-tuned four representative embedding models (`bge-large-en-v1.5`, `multilingual-e5-large`, `e5-large-v2`, and `qwen3-embedding-0.6B`) on our training split. Figure 3 visualizes the training dynamics for S@5 and EER. Due to space constraints, fine-grained results are reported in Appendix B.

Throughout the training process, all models exhibit significant performance gains, improving by approximately 20 percentage points on S@5 and reducing EER by 0.15. Most notably, the fine-tuned `multilingual-e5-large` establishes a new SOTA on the leaderboard, surpassing the previous best model (`llama3.1-8b`) by roughly 10 points in S@5 and 0.04 in EER. These results indicate that while current SOTA models are powerful generalized feature extractors, they still lack adaptation for specific downstream tasks like authorship attribution. AuthBench serves as an effective resource for bridging this gap through post-training.

6 Conclusion

In this work, we introduced **AuthBench**, a comprehensive benchmark designed to standardize the evaluation of authorship representation across diverse languages, genres, and document lengths. Our extensive evaluation demonstrates that while base LLMs currently provide the strongest stylistic signals—significantly outperforming traditional embedding models in complex scenarios—the task remains far from solved. Significant performance gaps persist, particularly in short context attribution and cross-genre generalization. Furthermore, our training experiments confirm that AuthBench serves as an effective substrate for domain adaptation, enabling models to better capture subtle authorship cues. We hope this resource facilitates future research into more robust, efficient, and universally adaptable authorship analysis.

Limitations

A primary limitation of this work is the uneven distribution of document lengths and genres across languages, which stems from the inherent characteristics of the public corpora utilized. We aim to address this distributional imbalance in future iterations by curating more balanced sources. Additionally, due to the large scale of AuthBench, manual

verification was feasible only for a representative subset of the data. However, the robust training dynamics and performance gains demonstrated in Section 5.5 serve as strong empirical evidence of the dataset's high quality and effective labeling. In future work, we plan to further inspect all the examples in the benchmark.

References

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

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*.

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. 2024a. Authorship attribution in the era of LLMs: Problems, methodologies, and challenges. *ACM SIGKDD Explorations*. ArXiv preprint arXiv:2408.08946.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024b. Can large language models identify authorship? *Findings of the Association for Computational Linguistics: EMNLP 2024*. ArXiv:2403.08213.

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.

Fereshteh Jafariakinabad and Kien A. Hua. 2019. Style-aware neural model with application in authorship attribution. In *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 325–328. IEEE.

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.

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.

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.

Panagiotis D. Michailidis, Ilias Papastamatiou, and Georgios Tzimas. 2022. A scientometric study of the stylometric research field. *Informatics*, 9(3):60.

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).

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)*.

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.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 363–373.

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.

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

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.

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.

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 normalization, deduplication, safety and quality filtering heuristics, bucketed sampling, and split construction. Section 3.1 in the main paper contains a concise summary.

AuthBench is constructed via a unified, reproducible pipeline that transforms heterogeneous raw corpora into a balanced and quality-controlled benchmark. The pipeline enforces (i) schema consistency across sources, (ii) strict de-duplication to reduce leakage and redundancy, (iii) conservative safety and quality filtering, and (iv) stratified split construction that preserves distributional balance across key axes.

A.1 Source Ingestion and Normalization

Let $\mathcal{S} = \{S_1, \ldots, S_m\}$ denote the set of raw sources (platforms, domains, and languages). Each source yields a multiset of candidate items $\mathcal{D}_i = \{d_{i,1}, \ldots, d_{i,n_i}\}$ with raw text and source-specific metadata. We standardize all items into a unified record

$$x = (\texttt{id},\ \ell,\ g,\ \sigma,\ \tau,\ t,\ \mathbf{y}),$$

where ℓ is language, g is (macro-)genre/domain, σ is the source identifier, τ is the task type (if applicable), t is the text field used for retrieval/attribution, and $\mathbf{y}$ denotes optional structured labels (e.g., class label, reference answer).

We apply deterministic text normalization $N(\cdot)$ to each raw text string: (i) Unicode canonicalization to NFC, (ii) removal of control characters, and (iii) whitespace canonicalization (collapse repeated whitespace; trim leading/trailing spaces). We tokenize with a fixed tokenizer $T(\cdot)$ used throughout the benchmark and compute token length

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

We retain minimal provenance metadata (language, source, genre/domain, task type) and store $L(t)$ for downstream filtering and bucketing.

A.2 De-duplication and Near-Duplicate Removal

To reduce redundancy and prevent train–test contamination via repeated content, we perform exact and near-duplicate removal within the candidate pool prior to splitting.

Exact duplicates. For each item with normalized text t, we compute a stable hash $h(t)$ (e.g., SHA-256) and drop duplicates by keeping a single representative per unique hash:

$$\mathcal{D}^{\text{exact}} = \{x \in \cup_i \mathcal{D}_i \ :\ h(t) \text{ is unique}\}.$$

Near duplicates. We represent each text t using character n-grams (we use $n = 5$) and apply a locality-sensitive hashing (LSH) scheme (e.g., MinHash) to approximate Jaccard similarity between texts. For two texts t and t', let $G(t)$ be the multiset of 5-grams and define

$$J(t, t') = \frac{|G(t) \cap G(t')|}{|G(t) \cup G(t')|}.$$

Given a threshold θ_{nd}, we remove x if there exists a previously accepted item x' such that $J(t, t') \geq \theta_{\text{nd}}$. We additionally remove common boilerplate patterns via deterministic rules (e.g., copyright headers, navigation lists) before similarity computation to avoid spurious matches.

A.3 Safety and Quality Filtering

We apply conservative safety and quality filters to exclude unsafe, malformed, or low-information items. An item is retained if it passes all filters below.

Policy and toxicity. Let $s_{\text{tox}}(t) \in [0, 1]$ denote a toxicity/policy risk score produced by one or more classifiers. We drop items above a conservative threshold τ_{tox}:

$$s_{\text{tox}}(t) > \tau_{\text{tox}} \Rightarrow \text{drop}.$$

Language identification and script consistency. Let $\hat{\ell}(t)$ be the predicted language (optional) and let S_ℓ be the expected Unicode script set for language ℓ (e.g., Latin for `en/es/fr/de`, Cyrillic for `ru`, Arabic for `ar`, CJK scripts for `zh/ja/ko`). Define the script-match ratio over letters:

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

evaluated when t contains at least 8 letters. We drop if $r_{\text{script}}(t;\ell) < \tau_{\text{script}}(\ell)$. If r_{script} is low/undefined and language identification is enabled, we additionally drop when $\hat{\ell}(t)$ is incompatible with the target language ℓ.

Length bounds. We enforce per-task token limits to avoid trivially short or excessively long instances. For each task type τ, we specify bounds $(L_{\min}(\tau), L_{\max}(\tau))$ and drop if

$$L(t) < L_{\min}(\tau) \quad \text{or} \quad L(t) > L_{\max}(\tau).$$

We also discard empty items where $L(t) = 0$.

Low-information and formatting heuristics. We compute a set of interpretable text-quality statistics and filter via thresholds. Let $w = T(t)$ be the token sequence.

- **Unique token ratio:**

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

  drop if $r_{\text{uniq}}(t) < \tau_{\text{uniq}}$.

- **Symbol ratio:** let $\#\text{sym}(t)$ count non-letter and non-digit characters (whitespace handling is source-dependent). Define

$$r_{\text{sym}}(t) \;=\; \frac{\#\text{sym}(t)}{|t|_{\text{chars}}},$$

  drop if $r_{\text{sym}}(t) > \tau_{\text{sym}}$.

- **Maximum repeated-character run:** let $m_{\text{rep}}(t)$ be the maximum run length of identical non-space characters. Drop if $m_{\text{rep}}(t) > K_{\text{rep}}$.

- **Alphabetic character ratio:**

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

  drop if $r_\alpha(t) < \tau_\alpha$.

- **Alphabetic token ratio:** let $\mathbb{I}[\cdot]$ denote the indicator function, and define

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

  where $\alpha(w_j)$ is true if token w_j contains at least one alphabetic character. We drop an item if $r_{\text{tok}}(t) < \tau_{\text{tok}}$.

- **Single-letter density:** let u_j be true when w_j is a single-letter token. Define

$$r_1(t) = \frac{1}{|w|}\sum_{j=1}^{|w|}\mathbb{I}[u_j],$$
$$m_1(t) = \max\{\text{length of any consecutive run of } u_j\}.$$

  We drop an item if $r_1(t) > \tau_1$ or $m_1(t) \geq K_1$.

For public-domain style corpora where heavy punctuation is common, we optionally include a maximum consecutive-symbol constraint $m_{\text{sym}}(t) \leq K_{\text{sym}}$.

A.4 Task-Specific Structuring and Validation

Some sources provide structured fields (e.g., question–answer pairs, labels, prompts, references). For each task type τ, we define a schema $\mathcal{F}_\tau$ (required fields) and retain an instance only if all required fields are present and non-empty after normalization:

$$\forall f \in \mathcal{F}_\tau,\;\; \texttt{field}_f \neq \varnothing.$$

Malformed entries are discarded. For long documents exceeding task-specific caps, we segment into contiguous chunks that preserve local coherence; each chunk inherits provenance metadata and receives a unique `id` derived from the original record and chunk offset.

A.5 Balanced Bucketed Sampling

The filtered pool can be imbalanced across languages, lengths, and domains. We therefore construct a balanced benchmark via bucketed sampling over salient axes.

Let each item x be assigned a bucket key

$$b(x) \;=\; (\ell,\; k,\; g,\; \tau),$$

where ℓ is language, τ is task type, g is genre/-domain, and k is a token-length bin derived from

$L(t)$. Length bins may be fixed ranges or quantiles computed per language. For each bucket b, let N_b be its available count.

We specify target proportions (or quotas) Q_b that enforce: (i) approximately balanced mass across languages, (ii) balanced length bins within each language, and (iii) bounded dominance of any single domain/task. We then sample uniformly without replacement within each bucket:

$$\mathcal{B} \;=\; \bigcup_{b} \text{Sample}\big(\{x : b(x) = b\},\; Q_b\big).$$

If some buckets are underfull ($N_b < Q_b$), we redistribute the deficit to sibling buckets that share (ℓ, τ) while preserving global balance constraints.

To scale sampling without materializing the full dataset in memory, we implement a streaming variant: we maintain per-bucket reservoirs and insert items in a single pass, approximating global uniform sampling subject to bucket constraints.

A.6 Post-processing and Final De-duplication

We apply a final canonicalization pass to standardize formatting (trim, normalize quotes, enforce a single trailing newline). We re-tokenize after canonicalization and re-check length bounds to ensure that trimming does not move instances outside $(L_{\min}, L_{\max})$.

To further mitigate leakage in prompt-based tasks, we perform a final de-duplication pass on selected fields (e.g., prompts or questions) by hashing normalized prompt text and ensuring uniqueness within the benchmark.

A.7 Split Construction

We construct splits (train/dev/test or dev/test, depending on the benchmark protocol) via stratification over the same bucket space to preserve balance. For each item x, we compute a split assignment using a deterministic hash of its identifier:

$$u(x) \;=\; \text{Hash}(\texttt{id}) \bmod 1,$$

and map $u(x)$ to split labels according to target split proportions, subject to per-bucket constraints. We enforce disjointness by design: items with identical or near-duplicate content are never placed in different splits, and when sources require provenance isolation, we additionally enforce source-level disjointness (no cross-split collisions within the specified unit, e.g., document or thread).

Overall, this pipeline yields a diverse benchmark with controlled coverage over language and length, reduced redundancy, conservative safety guarantees, and splits that preserve stratified balance for reliable evaluation.

A.8 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, reporting EER. 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).

A.9 Metrics

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

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

For $K = 10$, we compute:

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

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

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

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

We define Success@K as:

$$\text{Success@}K(q) = \begin{cases} 1 & \text{if } \sum_{i=1}^{K} \text{rel}_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We report $S@5$, $R@5$, and nDCG@5 as macro-averages 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

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

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

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

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

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

A.10 Models Evaluated

B Full Results Tables

Table 8 to 14 are the complete results across all models

B.1 Overall Leaderboard Full Results

Table 8 is the full results of overall leaderboard.

B.2 Overall Leaderboard Full Results

B.3 Language-wise Full Results

Table 9 and Table 10 are the complete finegrain results of each language for all models.

B.4 Primar-genre Full Results

Table 11 and Table 12 are the complete finegrain results of each language for all models.

B.5 Length-bucket Full Results

Table 13 and Table 14 are the complete finegrain results of each language for all models.

C Training Dynamic of Loss

Figure 4 illustrates the training loss dynamics of selected embedding models on AuthBench. All models exhibit a smooth and monotonic decrease in loss, indicating stable optimization and effective learning. Despite differences in convergence rates and initial loss values, the consistent downward trends suggest that AuthBench provides a meaningful training signal that can be reliably leveraged by diverse model architectures.

D Additional Dataset Statistics

Figure 5 to 8 shows the visualization of evaluation results across all models.

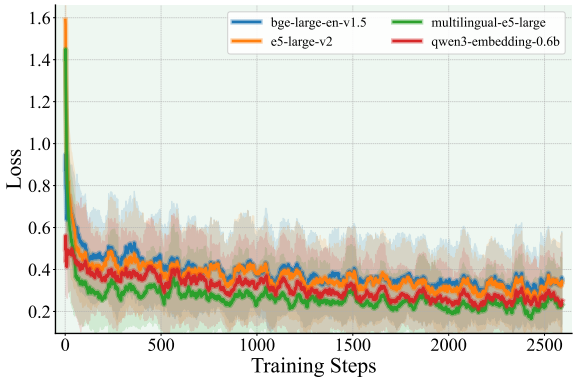

Figure 4: Training dynamics of loss of select models.

E Raw Data Sources

AuthBench consolidates 13 publicly available corpora 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. We focus on sources that provide author identifiers or (in limited cases) consistent author-like groupings; see Table 15.

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.

F Responsible NLP Checklist Details

This appendix provides checklist-specific details for ACL submission items B4, B6, C2, C3, D1, D3, D4, and E1. Where applicable, we point to the primary sections in the main paper and provide additional operational details here.

F.1 Personally Identifying Information (PII) or Offensive Content

PII. AuthBench is constructed exclusively from publicly available datasets with established licenses and distribution practices (Appendix E). We do not release raw author identifiers. Instead, we anonymize author identifiers by hashing (Section 3; Section 3.1). We also retain only minimal provenance metadata needed for analysis and reproducibility (language, genre, source, token length), and we do not include user profile fields,

Model	Hugging Face Repository
e5-large-v2	intfloat/e5-large-v2
e5-base-v2	intfloat/e5-base-v2
e5-small-v2	intfloat/e5-small-v2
e5-mistral-7b-instruct	intfloat/e5-mistral-7b-instruct
instructor-xl	hkunlp/instructor-xl
instructor-large	hkunlp/instructor-large
instructor-base	hkunlp/instructor-base
multilingual-e5-large	intfloat/multilingual-e5-large
multilingual-e5-base	intfloat/multilingual-e5-base
bge-large-en-v1.5	BAAI/bge-large-en-v1.5
bge-base-en-v1.5	BAAI/bge-base-en-v1.5
bge-small-en-v1.5	BAAI/bge-small-en-v1.5
bge-large-zh-v1.5	BAAI/bge-large-zh-v1.5
bge-base-zh-v1.5	BAAI/bge-base-zh-v1.5
bge-small-zh-v1.5	BAAI/bge-small-zh-v1.5
bge-m3	BAAI/bge-m3
snowflake-arctic-embed-l-v2	Snowflake/snowflake-arctic-embed-l-v2.0
snowflake-arctic-embed-m-v2	Snowflake/snowflake-arctic-embed-m-v2.0
jina-embeddings-v2-base-en	jinaai/jina-embeddings-v2-base-en
jina-embeddings-v2-small-en	jinaai/jina-embeddings-v2-small-en
mxbai-embed-large-v1	mixedbread-ai/mxbai-embed-large-v1
gte-large-en-v1.5	Alibaba-NLP/gte-large-en-v1.5
gte-qwen2-7b-instruct	Alibaba-NLP/gte-Qwen2-7B-instruct
gte-base	thenlper/gte-base
gte-large	thenlper/gte-large
nv-embed-v1	nvidia/NV-Embed-v1
qwen3-embedding-0.6b	Qwen/Qwen3-Embedding-0.6B
qwen3-embedding-4b	Qwen/Qwen3-Embedding-4B
qwen3-embedding-8b	Qwen/Qwen3-Embedding-8B
llama3-8b	meta-llama/Meta-Llama-3-8B
llama3-8b-instruct	meta-llama/Meta-Llama-3-8B-Instruct
llama3.1-8b	meta-llama/Llama-3.1-8B
llama3.1-8b-instruct	meta-llama/Llama-3.1-8B-Instruct
llama2-7b	meta-llama/Llama-2-7b-hf
llama2-7b-chat	meta-llama/Llama-2-7b-chat-hf
deepseek-llm-7b-base	deepseek-ai/deepseek-llm-7b-base
deepseek-llm-7b-chat	deepseek-ai/deepseek-llm-7b-chat
deepseek-coder-6.7b-instruct	deepseek-ai/deepseek-coder-6.7b-instruct

Table 7: **Models evaluated in AuthBench.** We list all embedding models and LLMs evaluated in this work together with their corresponding Hugging Face repositories.

URLs, or platform-specific identifiers beyond the hashed author ID.

Offensive or unsafe content. We apply conservative safety filtering during construction (Section 3.1; Appendix A.3). Specifically, we filter items using policy/toxicity classifiers (dropping items above a conservative threshold), enforce language and script consistency checks, and remove malformed or low-information items using interpretable heuristics (Appendix A.3). We additionally deduplicate and near-deduplicate content to reduce leakage and repeated unsafe fragments (Appendix A.2).

Residual risk. Because sources originate from the open web and large public corpora, residual PII or offensive text may persist despite filtering and deduplication. We mitigate this risk through anonymization, conservative filtering, provenance

Model	Size	S@5 ↑	R@5 ↑	nDCG@5 ↑	EER ↓
LLMs (instruction-tuned)					
deepseek-coder-6.7b-instruct	6.7B	0.455	0.450	0.368	0.095
deepseek-llm-7b-chat	7B	0.446	0.439	0.354	0.107
llama3-8b-instruct	8B	**0.530**	**0.524**	**0.428**	0.086
llama3.1-8b-instruct	8B	0.527	0.520	0.427	0.085
qwen2.5-3b-instruct	3.1B	0.498	0.491	0.396	**0.078**
qwen2.5-7b-instruct	7.6B	0.487	0.481	0.390	0.081
qwen3-4b-instruct	4B	0.483	0.476	0.381	0.089
LLMs (base)					
deepseek-llm-7b-base	7B	0.479	0.473	0.379	0.095
llama3-8b	8B	**0.542**	**0.537**	**0.441**	0.089
llama3.1-8b	8B	0.534	0.528	0.436	0.089
qwen2.5-3b	3.1B	0.496	0.488	0.394	**0.081**
qwen3-4b	4B	0.481	0.475	0.385	0.089
Embedding models (instruction-tuned)					
e5-mistral-7b-instruct	7.1B	0.467	0.460	0.384	0.099
gte-qwen2-7b-instruct	7.6B	**0.490**	**0.484**	**0.398**	**0.089**
sfr-embedding-mistral	7.1B	0.468	0.462	0.383	0.099
Embedding models					
all-minilm-l12-v2	33M	0.322	0.318	0.254	0.161
all-minilm-l6-v2	23M	0.322	0.318	0.251	0.149
all-mpnet-base-v2	109M	0.346	0.340	0.272	0.144
all-roberta-large-v1	355M	0.380	0.374	0.296	0.127
allenai-specter	110M	0.312	0.310	0.251	0.152
bert-base-uncased	110M	0.394	0.388	0.311	**0.103**
bge-base-en-v1.5	109M	0.360	0.355	0.282	0.172
bge-base-zh-v1.5	102M	0.369	0.365	0.297	0.150
bge-large-en-v1.5	335M	0.357	0.352	0.284	0.163
bge-large-zh-v1.5	326M	0.372	0.367	0.294	0.167
bge-m3	568M	0.326	0.322	0.263	0.272
bge-small-en-v1.5	33M	0.329	0.326	0.264	0.190
distiluse-base-multilingual-cased-v2	135M	0.329	0.323	0.266	0.220
e5-base-v2	109M	0.412	0.406	0.328	0.170
e5-large-v2	335M	0.393	0.388	0.319	0.181
e5-small-v2	33M	0.358	0.354	0.293	0.211
facebook-contriever	109M	0.443	0.439	0.367	0.125
facebook-contriever-msmarco	109M	0.368	0.365	0.294	0.175
gte-base	n/a	0.352	0.348	0.274	0.153
gte-large	335M	0.355	0.351	0.285	0.134
gte-large-en-v1.5	434M	0.394	0.389	0.314	0.120
jina-embeddings-v2-base-en	137M	0.111	0.110	0.086	0.303
jina-embeddings-v2-small-en	33M	0.311	0.308	0.242	0.112
msmarco-distilbert-base-v4	66M	0.324	0.320	0.255	0.227
multilingual-e5-base	278M	0.458	0.454	0.373	0.155
multilingual-e5-large	560M	**0.485**	**0.479**	**0.398**	0.145
mxbai-embed-large-v1	335M	0.353	0.348	0.278	0.165
nomic-embed-text-v1	n/a	0.370	0.367	0.295	0.154
nomic-embed-text-v1.5	137M	0.370	0.366	0.295	0.132
paraphrase-mpnet-base-v2	n/a	0.344	0.336	0.271	0.144
paraphrase-multilingual-mpnet-base-v2	278M	0.284	0.282	0.232	0.256
qwen3-embedding-0.6b	596M	0.421	0.415	0.334	0.113
qwen3-embedding-4b	4B	0.438	0.434	0.354	0.107
qwen3-embedding-8b	7.6B	0.468	0.463	0.371	0.127
snowflake-arctic-embed-l-v2	568M	0.355	0.351	0.289	0.202
Lexical baseline					
tfidf	n/a	**0.326**	**0.321**	**0.264**	**0.177**

Table 8: **Overall results on AuthBench.** Authorship attribution is evaluated with Success@5 (S@5), Recall@5 (R@5), and nDCG@5 (higher is better). Authorship verification is evaluated with equal error rate (EER; lower is better). **Bold** = best, underlined = second best within each model group. **Size** denotes parameter count when publicly available.

Model	Size	ar	de	en	es	fr	hi	ja	ko	ru	zh
LLMs (instruction-tuned)											
deepseek-coder-6.7b-instruct	6.7B	0.305	0.662	0.410	0.535	0.581	0.525	**0.941**	0.435	0.391	0.504
deepseek-llm-7b-chat	7B	0.266	0.620	0.500	0.564	0.516	0.575	**0.941**	0.402	0.323	0.488
llama3-8b-instruct	8B	0.305	**0.761**	**0.566**	0.644	0.629	0.800	0.765	0.500	**0.504**	0.504
llama3.1-8b-instruct	8B	0.310	0.746	0.530	**0.693**	**0.645**	**0.850**	0.647	**0.511**	0.474	0.504
qwen2.5-3b-instruct	3.1B	**0.350**	0.648	0.464	0.594	0.629	0.600	0.765	0.478	0.459	**0.543**
qwen2.5-7b-instruct	7.6B	0.335	0.648	0.458	0.574	0.597	0.650	0.882	0.424	0.444	**0.543**
qwen3-4b-instruct	4B	0.340	0.662	0.434	0.564	0.581	0.700	0.765	0.457	0.421	**0.543**
LLMs (base)											
deepseek-llm-7b-base	7B	0.266	0.648	0.524	0.634	0.613	0.675	**0.941**	0.424	0.361	0.520
llama3-8b	8B	0.335	**0.775**	**0.560**	0.663	**0.645**	**0.850**	0.824	**0.500**	**0.496**	0.512
llama3.1-8b	8B	0.335	**0.775**	0.554	**0.673**	**0.645**	0.800	0.765	0.478	0.489	0.496
qwen2.5-3b	3.1B	**0.345**	0.634	0.458	0.594	0.629	0.600	0.882	0.489	0.444	**0.543**
qwen3-4b	4B	0.340	0.676	0.464	0.574	0.565	0.625	0.706	0.446	0.406	0.535
Embedding models (instruction-tuned)											
e5-mistral-7b-instruct	7.1B	0.276	**0.648**	0.476	**0.604**	0.532	0.625	**0.882**	**0.457**	0.451	0.441
gte-qwen2-7b-instruct	7.6B	**0.296**	0.634	**0.488**	0.564	**0.645**	**0.675**	0.765	**0.457**	**0.474**	**0.535**
sfr-embedding-mistral	7.1B	0.271	0.634	**0.488**	**0.604**	0.548	0.650	0.706	**0.457**	0.444	0.465
Embedding models											
all-minilm-l12-v2	33M	0.163	0.380	0.386	0.396	0.387	0.600	0.471	0.326	0.248	0.339
all-minilm-l6-v2	23M	0.192	0.408	0.380	0.436	0.306	0.475	0.294	0.293	0.256	0.370
all-mpnet-base-v2	109M	0.192	0.507	0.470	0.406	0.323	0.525	0.353	0.283	0.248	0.394
all-roberta-large-v1	355M	0.266	0.634	0.428	0.465	0.419	0.575	0.647	0.283	0.263	0.370
allenai-specter	110M	0.167	0.577	0.343	0.455	0.403	0.325	0.294	0.326	0.241	0.260
bert-base-uncased	110M	0.212	0.563	0.458	0.495	0.403	0.575	0.588	0.424	0.316	0.402
bge-base-en-v1.5	109M	0.197	0.592	0.386	0.455	0.371	0.625	0.588	0.348	0.286	0.346
bge-base-zh-v1.5	102M	0.212	0.535	0.349	0.406	0.323	0.650	**0.824**	0.424	0.278	0.449
bge-large-en-v1.5	335M	0.153	0.620	0.428	0.475	0.452	0.525	0.353	0.326	0.271	0.362
bge-large-zh-v1.5	326M	0.241	0.521	0.325	0.406	0.323	0.625	0.529	0.424	0.323	0.465
bge-m3	568M	0.232	0.437	0.319	0.505	0.339	0.400	0.176	0.261	0.263	0.386
bge-small-en-v1.5	33M	0.148	0.493	0.446	0.426	0.387	0.500	0.294	0.304	0.226	0.346
distiluse-base-multilingual-cased-v2	135M	0.212	0.479	0.349	0.426	0.274	0.500	0.353	0.380	0.226	0.370
e5-base-v2	109M	0.256	0.563	0.476	0.545	0.468	**0.750**	0.647	0.348	0.301	0.386
e5-large-v2	335M	0.227	0.521	0.464	0.475	0.516	0.600	0.529	0.391	0.301	0.386
e5-small-v2	33M	0.212	0.465	0.428	0.396	0.371	0.650	0.647	0.326	0.308	0.346
facebook-contriever	109M	0.227	0.648	**0.536**	0.614	0.597	0.625	0.647	0.413	0.338	0.386
facebook-contriever-msmarco	109M	0.202	0.521	0.446	0.446	0.355	0.600	0.588	0.391	0.286	0.354
gte-base	n/a	0.143	0.592	0.446	0.485	0.419	0.500	0.588	0.272	0.278	0.346
gte-large	335M	0.143	0.620	0.416	0.535	0.484	0.500	0.353	0.304	0.271	0.339
gte-large-en-v1.5	434M	0.222	0.634	0.446	0.515	0.516	0.600	0.294	0.326	0.323	0.386
jina-embeddings-v2-base-en	137M	0.128	0.099	0.060	0.050	0.048	0.275	0.176	0.152	0.053	0.205
jina-embeddings-v2-small-en	33M	0.163	0.423	0.392	0.327	0.274	0.500	0.588	0.380	0.203	0.354
msmarco-distilbert-base-v4	66M	0.192	0.479	0.386	0.297	0.258	0.600	0.529	0.402	0.263	0.315
multilingual-e5-base	278M	0.310	0.620	0.506	0.545	0.532	0.625	0.412	0.391	0.383	0.512
multilingual-e5-large	560M	**0.335**	**0.676**	0.488	**0.644**	**0.629**	0.625	0.647	0.370	**0.459**	0.465
mxbai-embed-large-v1	335M	0.138	0.592	0.416	0.475	0.452	0.575	0.412	0.326	0.278	0.354
nomic-embed-text-v1	n/a	0.207	0.549	0.440	0.485	0.500	0.575	0.471	0.293	0.271	0.362
nomic-embed-text-v1.5	137M	0.192	0.549	0.446	0.525	0.484	0.550	0.529	0.272	0.278	0.362
paraphrase-mpnet-base-v2	n/a	0.227	0.493	0.428	0.376	0.306	0.525	0.471	0.348	0.263	0.339
paraphrase-multilingual-mpnet-base-v2	278M	0.192	0.366	0.355	0.356	0.177	0.425	0.176	0.261	0.241	0.315
qwen3-embedding-0.6b	596M	0.276	0.563	0.428	0.525	0.500	0.550	0.706	0.380	0.361	0.457
qwen3-embedding-4b	4B	0.251	0.606	0.464	0.564	0.597	0.600	0.647	0.391	0.331	0.496
qwen3-embedding-8b	7.6B	0.323	0.600	0.404	0.545	0.581	0.600	0.706	**0.467**	0.429	**0.543**
snowflake-arctic-embed-l-v2	568M	0.232	0.423	0.373	0.505	0.290	0.550	0.059	0.348	0.316	0.425
Lexical baseline											
tfidf	n/a	**0.291**	**0.310**	**0.313**	**0.267**	**0.419**	**0.575**	**0.529**	**0.380**	**0.293**	**0.299**

Table 9: **Language-wise Success@5 on AuthBench (full results).** Authorship attribution is evaluated with Success@5 (S@5), Recall@5 (R@5), and nDCG@5 (higher is better). Authorship verification is evaluated with equal error rate (EER; lower is better). **Bold** = best, underlined = second best within each model group. **Size** denotes parameter count when publicly available.

Model	Size	ar	de	en	es	fr	hi	ja	ko	ru	zh
LLMs (instruction-tuned)											
deepseek-coder-6.7b-instruct	6.7B	0.108	0.057	0.110	0.075	0.062	0.041	0.012	0.074	0.089	0.068
deepseek-llm-7b-chat	7B	0.130	0.095	0.111	0.084	0.065	0.049	0.009	0.087	0.080	0.077
llama3-8b-instruct	8B	0.087	**0.042**	**0.076**	0.065	**0.041**	**0.025**	0.011	**0.052**	0.063	0.065
llama3.1-8b-instruct	8B	**0.080**	**0.042**	0.077	0.065	0.046	**0.025**	0.013	0.053	0.066	0.067
qwen2.5-3b-instruct	3.1B	0.082	0.056	0.097	**0.060**	0.048	0.043	0.013	0.068	**0.062**	0.066
qwen2.5-7b-instruct	7.6B	0.082	0.043	0.082	0.063	0.050	0.036	**0.008**	**0.052**	0.073	0.068
qwen3-4b-instruct	4B	0.092	0.052	0.092	0.062	0.057	0.035	0.015	0.064	0.069	**0.062**
LLMs (base)											
deepseek-llm-7b-base	7B	0.103	0.050	0.105	0.075	0.054	0.036	0.014	0.069	0.075	0.074
llama3-8b	8B	0.085	0.042	0.069	**0.061**	**0.040**	**0.025**	**0.011**	0.046	0.066	**0.064**
llama3.1-8b	8B	**0.083**	**0.039**	**0.066**	0.064	**0.040**	**0.025**	**0.011**	**0.045**	**0.058**	0.068
qwen2.5-3b	3.1B	0.088	0.056	0.096	**0.061**	0.050	0.035	0.012	0.062	0.061	0.066
qwen3-4b	4B	0.097	0.042	0.089	0.064	0.053	0.027	0.015	0.072	0.068	0.068
Embedding models (instruction-tuned)											
e5-mistral-7b-instruct	7.1B	0.103	**0.042**	0.086	0.071	**0.046**	0.038	0.016	**0.063**	0.068	0.070
gte-qwen2-7b-instruct	7.6B	0.111	**0.042**	**0.071**	**0.065**	**0.046**	**0.029**	**0.012**	**0.063**	**0.058**	**0.060**
sfr-embedding-mistral	7.1B	**0.099**	**0.042**	0.083	0.072	**0.046**	0.040	0.018	0.067	0.072	0.070
Embedding models											
all-minilm-l12-v2	33M	0.148	0.194	0.232	0.075	0.102	0.072	0.111	0.095	0.095	0.113
all-minilm-l6-v2	23M	0.152	0.151	0.249	0.084	0.087	0.071	0.078	0.097	0.075	0.114
all-mpnet-base-v2	109M	0.139	0.105	0.230	0.089	0.118	0.050	0.111	0.083	0.078	0.120
all-roberta-large-v1	355M	0.117	0.086	0.212	0.070	0.077	0.051	0.111	0.072	0.086	0.098
allenai-specter	110M	0.209	0.069	0.194	0.093	0.108	0.175	0.139	0.124	0.117	0.171
bert-base-uncased	110M	0.126	0.056	0.136	0.076	0.062	0.040	0.056	0.076	0.086	0.105
bge-base-en-v1.5	109M	0.135	0.111	0.331	0.103	0.168	0.038	0.084	0.098	0.091	0.143
bge-base-zh-v1.5	102M	0.138	0.058	0.194	0.084	0.123	0.075	0.018	0.082	0.102	0.188
bge-large-en-v1.5	335M	0.133	0.097	0.296	0.129	0.138	0.037	0.048	0.083	0.108	0.113
bge-large-zh-v1.5	326M	0.160	0.097	0.269	0.109	0.154	0.100	0.111	0.113	0.114	0.248
bge-m3	568M	0.246	0.249	0.311	0.168	0.231	0.300	0.333	0.316	0.321	0.263
bge-small-en-v1.5	33M	0.161	0.097	0.237	0.131	0.118	0.050	0.056	0.156	0.098	0.143
distiluse-base-multilingual-cased-v2	135M	0.185	0.194	0.208	0.170	0.246	0.189	0.167	0.258	0.270	0.256
e5-base-v2	109M	0.112	0.112	0.243	0.108	0.149	0.074	0.048	0.105	0.131	0.135
e5-large-v2	335M	0.139	0.161	0.207	0.125	0.154	0.059	0.111	0.113	0.182	0.165
e5-small-v2	33M	0.171	0.194	0.295	0.169	0.215	0.087	0.126	0.124	0.187	0.180
facebook-contriever	109M	0.107	0.071	0.154	0.066	0.092	0.033	0.056	0.080	**0.073**	0.099
facebook-contriever-msmarco	109M	0.143	0.125	0.249	0.112	0.184	0.052	0.064	0.093	0.095	0.110
gte-base	n/a	0.137	0.083	0.269	0.091	0.092	0.040	0.056	0.072	0.077	0.128
gte-large	335M	0.122	0.056	0.247	0.083	0.062	0.031	0.033	0.072	0.090	0.105
gte-large-en-v1.5	434M	0.117	0.055	0.260	0.075	0.053	0.040	0.081	0.109	0.096	0.113
jina-embeddings-v2-base-en	137M	0.166	0.345	0.444	0.290	0.324	0.318	0.206	0.296	0.295	0.195
jina-embeddings-v2-small-en	33M	0.121	0.083	0.147	0.112	0.131	**0.030**	0.056	0.086	0.091	0.094
msmarco-distilbert-base-v4	66M	0.155	0.097	0.243	0.084	0.118	0.037	0.162	0.134	0.115	0.113
multilingual-e5-base	278M	0.166	0.141	0.178	0.109	0.122	0.125	0.193	0.103	0.223	0.090
multilingual-e5-large	560M	0.148	0.139	0.160	0.099	0.106	0.080	0.111	0.125	0.201	**0.078**
mxbai-embed-large-v1	335M	0.136	0.110	0.299	0.112	0.134	0.034	0.056	0.077	0.102	0.113
nomic-embed-text-v1	n/a	0.117	0.059	0.237	0.068	0.081	0.043	0.060	0.113	0.077	0.135
nomic-embed-text-v1.5	137M	0.119	**0.054**	0.207	**0.065**	**0.046**	0.041	0.056	0.082	0.076	0.120
paraphrase-mpnet-base-v2	n/a	0.143	0.070	0.179	0.085	0.092	0.107	0.111	0.113	0.096	0.120
paraphrase-multilingual-mpnet-base-v2	278M	0.242	0.205	0.272	0.215	0.262	0.227	0.278	0.299	0.299	0.235
qwen3-embedding-0.6b	596M	0.139	0.109	**0.118**	0.103	0.054	0.099	0.009	**0.070**	0.079	0.098
qwen3-embedding-4b	4B	0.112	0.087	0.130	0.093	0.084	0.075	0.009	0.072	0.088	0.120
qwen3-embedding-8b	7.6B	**0.096**	0.106	0.162	0.123	0.121	0.039	**0.008**	0.078	0.095	0.091
snowflake-arctic-embed-l-v2	568M	0.164	0.208	0.219	0.150	0.246	0.104	0.333	0.229	0.241	0.188
Lexical baseline											
tfidf	n/a	**0.112**	**0.083**	**0.195**	**0.110**	**0.123**	**0.100**	**0.073**	**0.392**	**0.104**	**0.489**

Table 10: **Language-wise EER on AuthBench (full results).** Authorship attribution is evaluated with Success@5 (S@5), Recall@5 (R@5), and nDCG@5 (higher is better). Authorship verification is evaluated with equal error rate (EER; lower is better). **Bold** = best, underlined = second best within each model group. **Size** denotes parameter count when publicly available.

Model	Size	blog	ecommerce_reviews	literature	media_reviews	news	poetry	research_paper	social_media
LLMs (instruction-tuned)									
deepseek-coder-6.7b-instruct	6.7B	0.146	0.308	0.575	0.218	0.448	0.288	0.840	0.503
deepseek-llm-7b-chat	7B	0.174	0.333	0.511	0.164	0.462	0.308	**0.880**	0.514
llama3-8b-instruct	8B	**0.416**	0.385	**0.632**	0.200	**0.560**	0.288	**0.880**	0.592
llama3.1-8b-instruct	8B	0.412	0.359	**0.632**	0.200	0.538	0.308	**0.880**	**0.594**
qwen2.5-3b-instruct	3.1B	0.218	0.333	0.592	0.236	0.538	**0.327**	0.800	0.552
qwen2.5-7b-instruct	7.6B	0.156	**0.410**	0.586	0.236	0.513	0.288	0.760	0.585
qwen3-4b-instruct	4B	0.145	0.359	0.575	**0.255**	0.473	0.308	0.800	0.568
LLMs (base)									
deepseek-llm-7b-base	7B	0.323	0.385	0.569	0.200	0.487	0.250	**0.840**	0.540
llama3-8b	8B	**0.412**	0.385	0.626	0.182	0.574	**0.308**	**0.840**	**0.623**
llama3.1-8b	8B	0.351	0.333	**0.632**	0.182	**0.581**	**0.308**	**0.840**	0.600
qwen2.5-3b	3.1B	0.218	0.333	0.586	**0.255**	0.531	**0.308**	0.800	0.552
qwen3-4b	4B	0.273	**0.410**	0.575	**0.255**	0.502	**0.308**	0.800	0.530
Embedding models (instruction-tuned)									
e5-mistral-7b-instruct	7.1B	0.240	0.333	0.615	0.127	0.520	0.173	**0.840**	0.509
gte-qwen2-7b-instruct	7.6B	**0.251**	0.333	0.615	**0.236**	**0.538**	**0.250**	0.800	**0.542**
sfr-embedding-mistral	7.1B	0.246	**0.359**	**0.621**	0.164	0.527	0.154	**0.840**	0.503
Embedding models									
all-minilm-l12-v2	33M	0.200	0.128	0.391	0.109	0.256	0.058	0.880	0.407
all-minilm-l6-v2	23M	0.195	0.154	0.408	0.127	0.285	0.038	0.840	0.401
all-mpnet-base-v2	109M	**0.407**	0.256	0.414	0.164	0.296	0.096	0.840	0.379
all-roberta-large-v1	355M	0.207	0.333	0.437	0.164	0.310	**0.269**	0.840	0.416
allenai-specter	110M	0.218	0.154	0.466	0.055	0.292	0.058	0.800	0.351
bert-base-uncased	110M	0.241	0.282	0.540	0.127	0.397	0.115	0.800	0.409
bge-base-en-v1.5	109M	0.223	0.256	0.489	0.145	0.285	0.135	0.880	0.444
bge-base-zh-v1.5	102M	0.161	0.179	0.425	0.109	0.397	0.077	0.640	0.480
bge-large-en-v1.5	335M	0.278	0.256	0.506	0.164	0.282	0.115	0.840	0.405
bge-large-zh-v1.5	326M	0.134	0.231	0.437	0.182	0.401	0.154	0.600	0.433
bge-m3	568M	0.151	0.128	0.471	0.164	0.285	0.154	0.840	0.360
bge-small-en-v1.5	33M	0.340	0.308	0.448	0.109	0.271	0.096	0.880	0.363
distiluse-base-multilingual-cased-v2	135M	0.177	0.205	0.391	0.127	0.318	0.115	0.720	0.381
e5-base-v2	109M	0.300	0.385	0.511	0.164	0.390	0.173	0.840	0.457
e5-large-v2	335M	0.269	0.436	0.466	0.145	0.408	0.038	0.880	0.430
e5-small-v2	33M	0.273	0.308	0.460	0.109	0.321	0.038	0.800	0.401
facebook-contriever	109M	0.300	0.308	0.586	0.164	0.448	0.115	0.800	0.482
facebook-contriever-msmarco	109M	0.267	0.205	0.494	0.164	0.336	0.154	0.760	0.449
gte-base	n/a	0.261	0.256	0.540	0.164	0.289	0.038	**0.960**	0.381
gte-large	335M	0.261	0.205	0.517	0.145	0.296	0.115	0.920	0.393
gte-large-en-v1.5	434M	0.328	0.154	0.511	0.145	0.412	0.135	0.920	0.395
jina-embeddings-v2-base-en	137M	0.005	0.000	0.034	0.036	0.116	0.096	0.160	0.212
jina-embeddings-v2-small-en	33M	0.201	0.282	0.351	0.109	0.289	0.058	0.680	0.385
msmarco-distilbert-base-v4	66M	0.183	0.179	0.374	0.091	0.318	0.058	0.840	0.397
multilingual-e5-base	278M	0.277	**0.462**	0.517	0.218	0.487	0.173	0.840	0.497
multilingual-e5-large	560M	0.384	0.385	**0.626**	0.164	0.498	**0.269**	0.840	0.494
mxbai-embed-large-v1	335M	0.317	0.231	0.517	0.182	0.278	0.096	0.880	0.408
nomic-embed-text-v1	n/a	0.255	0.282	0.511	0.127	0.300	0.058	0.880	0.422
nomic-embed-text-v1.5	137M	0.328	0.359	0.511	0.145	0.282	0.077	0.840	0.423
paraphrase-mpnet-base-v2	n/a	0.351	0.308	0.397	0.145	0.285	0.096	0.800	0.403
paraphrase-multilingual-mpnet-base-v2	278M	0.184	0.256	0.356	0.091	0.235	0.077	0.680	0.320
qwen3-embedding-0.6b	596M	0.246	0.256	0.569	0.145	0.412	0.135	0.840	0.484
qwen3-embedding-4b	4B	0.313	0.333	0.592	0.164	0.455	0.115	0.840	0.509
qwen3-embedding-8b	7.6B	0.235	0.308	0.607	**0.236**	**0.513**	0.237	0.800	**0.518**
snowflake-arctic-embed-l-v2	568M	0.178	0.231	0.489	0.145	0.300	0.096	0.840	0.424
Lexical baseline									
tfidf	n/a	**0.172**	**0.103**	**0.379**	**0.073**	**0.310**	**0.250**	**0.800**	**0.385**

Table 11: **Primary-genre Success@5 on AuthBench (full results).** Authorship attribution is evaluated with Success@5 (S@5), Recall@5 (R@5), and nDCG@5 (higher is better). Authorship verification is evaluated with equal error rate (EER; lower is better). Values for `blog` and `social_media` are macro-averaged over fine-grained subgenres. **Bold** = best, underlined = second best within each model group. **Size** denotes parameter count when publicly available.

Model	Size	blog	ecommerce_reviews	literature	media_reviews	news	poetry	research_paper	social_media
LLMs (instruction-tuned)									
deepseek-coder-6.7b-instruct	6.7B	0.092	0.087	0.055	0.081	0.084	0.047	0.042	0.121
deepseek-llm-7b-chat	7B	0.080	0.067	0.060	0.086	0.098	0.078	0.068	0.127
llama3-8b-instruct	8B	0.073	**0.051**	**0.049**	0.079	0.087	0.040	**0.025**	**0.076**
llama3.1-8b-instruct	8B	0.077	0.053	**0.049**	0.080	0.084	**0.038**	0.026	0.078
qwen2.5-3b-instruct	3.1B	**0.059**	0.061	0.065	**0.071**	**0.061**	0.045	0.038	0.083
qwen2.5-7b-instruct	7.6B	0.064	0.055	0.060	0.080	0.077	0.048	0.026	0.080
qwen3-4b-instruct	4B	0.065	0.072	0.060	0.075	0.082	0.041	0.077	0.096
LLMs (base)									
deepseek-llm-7b-base	7B	0.091	0.061	0.065	0.081	0.089	0.062	0.077	0.109
llama3-8b	8B	0.070	**0.051**	**0.049**	**0.078**	0.086	0.042	0.023	0.074
llama3.1-8b	8B	0.074	**0.051**	**0.049**	0.083	0.086	**0.038**	**0.021**	**0.073**
qwen2.5-3b	3.1B	**0.059**	0.064	0.067	**0.078**	**0.065**	0.043	0.038	0.084
qwen3-4b	4B	0.064	0.053	0.060	0.080	0.083	0.043	0.043	0.105
Embedding models (instruction-tuned)									
e5-mistral-7b-instruct	7.1B	0.082	0.071	**0.062**	0.082	0.105	0.052	0.074	**0.093**
gte-qwen2-7b-instruct	7.6B	**0.060**	**0.066**	0.065	**0.071**	**0.090**	**0.043**	**0.018**	0.095
sfr-embedding-mistral	7.1B	0.076	0.068	0.063	0.084	0.105	0.051	0.072	**0.093**
Embedding models									
all-minilm-l12-v2	33M	0.242	0.333	0.109	0.152	0.162	0.153	0.077	0.151
all-minilm-l6-v2	23M	0.243	0.333	0.084	0.138	0.153	0.180	0.077	0.142
all-mpnet-base-v2	109M	0.232	0.285	0.087	0.131	0.153	0.130	0.115	0.143
all-roberta-large-v1	355M	0.221	0.129	0.084	0.093	0.144	0.130	0.038	0.102
allenai-specter	110M	0.135	0.205	0.067	0.182	0.147	0.159	0.077	0.186
bert-base-uncased	110M	**0.080**	0.092	0.071	0.119	0.105	0.118	0.094	0.136
bge-base-en-v1.5	109M	0.340	0.385	0.097	0.165	0.197	0.122	0.077	0.163
bge-base-zh-v1.5	102M	0.162	0.154	0.071	0.256	0.139	0.130	0.115	0.153
bge-large-en-v1.5	335M	0.329	0.359	0.097	0.141	0.177	0.102	0.038	0.149
bge-large-zh-v1.5	326M	0.174	0.256	0.075	0.306	0.133	0.174	0.154	0.164
bge-m3	568M	0.313	0.282	0.120	0.345	0.296	0.159	0.115	0.271
bge-small-en-v1.5	33M	0.239	0.256	0.114	0.175	0.213	0.116	0.038	0.171
distiluse-base-multilingual-cased-v2	135M	0.236	0.146	0.155	0.345	0.213	0.124	0.093	0.236
e5-base-v2	109M	0.201	0.205	0.120	0.168	0.191	0.105	0.091	0.165
e5-large-v2	335M	0.267	0.154	0.136	0.203	0.199	0.145	0.049	0.176
e5-small-v2	33M	0.308	0.205	0.147	0.207	0.251	0.200	0.115	0.198
facebook-contriever	109M	0.280	0.103	0.071	0.112	0.117	0.124	0.085	0.139
facebook-contriever-msmarco	109M	0.264	0.168	0.107	0.119	0.206	0.132	0.070	0.169
gte-base	n/a	0.270	0.359	0.083	0.138	0.174	0.135	0.059	0.140
gte-large	335M	0.226	0.365	0.076	0.127	0.132	0.116	0.077	0.125
gte-large-en-v1.5	434M	0.164	0.425	0.058	0.138	0.109	0.130	0.077	0.105
jina-embeddings-v2-base-en	137M	0.406	0.535	0.316	0.150	0.291	0.092	0.346	0.327
jina-embeddings-v2-small-en	33M	0.101	0.103	0.108	0.105	0.101	0.134	0.115	0.143
msmarco-distilbert-base-v4	66M	0.338	0.205	0.136	0.137	0.263	0.153	0.077	0.200
multilingual-e5-base	278M	0.181	0.128	0.109	0.099	0.134	0.106	0.077	0.133
multilingual-e5-large	560M	0.173	0.103	0.089	**0.086**	0.130	0.088	0.086	0.128
mxbai-embed-large-v1	335M	0.329	0.358	0.087	0.138	0.188	0.106	0.038	0.151
nomic-embed-text-v1	n/a	0.186	0.256	0.067	0.180	0.157	0.111	0.077	0.153
nomic-embed-text-v1.5	137M	0.199	0.205	0.065	0.155	0.128	0.113	0.077	0.140
paraphrase-mpnet-base-v2	n/a	0.177	0.103	0.077	0.130	0.157	0.113	0.091	0.155
paraphrase-multilingual-mpnet-base-v2	278M	0.323	0.205	0.179	0.224	0.307	0.132	0.077	0.261
qwen3-embedding-0.6b	596M	0.086	0.094	0.063	0.113	**0.096**	0.059	0.038	0.097
qwen3-embedding-4b	4B	0.090	**0.077**	**0.049**	0.155	0.101	0.058	0.073	0.096
qwen3-embedding-8b	7.6B	0.116	**0.077**	0.062	0.103	0.118	**0.039**	**0.000**	**0.088**
snowflake-arctic-embed-l-v2	568M	0.289	0.179	0.106	0.224	0.224	0.069	0.077	0.183
Lexical baseline									
tfidf	n/a	**0.138**	**0.179**	**0.103**	**0.531**	**0.153**	**0.087**	**0.077**	**0.179**

Table 12: **Primary-genre EER on AuthBench (full results).** Authorship attribution is evaluated with Success@5 (S@5), Recall@5 (R@5), and nDCG@5 (higher is better). Authorship verification is evaluated with equal error rate (EER; lower is better). Values for `blog` and `social_media` are macro-averaged over fine-grained subgenres. **Bold** = best, underlined = second best within each model group. **Size** denotes parameter count when publicly available.

Model	Size	short	medium	long	extra_long
LLMs (instruction-tuned)					
deepseek-coder-6.7b-instruct	6.7B	0.333	0.571	0.402	0.267
deepseek-llm-7b-chat	7B	0.370	0.532	0.403	0.320
llama3-8b-instruct	8B	0.444	0.607	**0.491**	**0.427**
llama3.1-8b-instruct	8B	**0.481**	**0.610**	0.489	0.373
qwen2.5-3b-instruct	3.1B	0.370	0.589	0.461	0.333
qwen2.5-7b-instruct	7.6B	0.444	0.563	0.459	0.307
qwen3-4b-instruct	4B	0.370	0.592	0.434	0.307
LLMs (base)					
deepseek-llm-7b-base	7B	0.370	0.571	0.442	0.307
llama3-8b	8B	**0.407**	0.623	**0.511**	**0.387**
llama3.1-8b	8B	**0.407**	**0.625**	0.493	**0.387**
qwen2.5-3b	3.1B	**0.407**	0.579	0.463	0.333
qwen3-4b	4B	**0.407**	0.584	0.432	0.320
Embedding models (instruction-tuned)					
e5-mistral-7b-instruct	7.1B	**0.370**	0.571	0.417	0.320
gte-qwen2-7b-instruct	7.6B	0.333	**0.581**	**0.455**	0.320
sfr-embedding-mistral	7.1B	**0.370**	0.579	0.411	**0.333**
Embedding models					
all-minilm-l12-v2	33M	0.370	0.344	0.300	0.347
all-minilm-l6-v2	23M	0.333	0.354	0.296	0.333
all-mpnet-base-v2	109M	0.444	0.351	0.327	**0.413**
all-roberta-large-v1	355M	0.444	0.398	0.371	0.333
allenai-specter	110M	0.296	0.398	0.256	0.267
bert-base-uncased	110M	0.333	0.494	0.333	0.333
bge-base-en-v1.5	109M	0.407	0.437	0.314	0.267
bge-base-zh-v1.5	102M	0.370	0.434	0.325	0.333
bge-large-en-v1.5	335M	0.370	0.429	0.312	0.293
bge-large-zh-v1.5	326M	0.370	0.434	0.319	**0.413**
bge-m3	568M	0.407	0.432	0.256	0.240
bge-small-en-v1.5	33M	0.407	0.359	0.308	0.293
distiluse-base-multilingual-cased-v2	135M	0.370	0.390	0.294	0.240
e5-base-v2	109M	0.370	0.460	0.382	0.387
e5-large-v2	335M	0.333	0.457	0.352	0.373
e5-small-v2	33M	0.370	0.403	0.331	0.307
facebook-contriever	109M	0.370	0.496	0.421	0.347
facebook-contriever-msmarco	109M	0.370	0.437	0.327	0.293
gte-base	n/a	0.370	0.408	0.321	0.267
gte-large	335M	0.407	0.424	0.308	0.307
gte-large-en-v1.5	434M	**0.481**	0.457	0.358	0.293
jina-embeddings-v2-base-en	137M	0.111	0.109	0.101	0.187
jina-embeddings-v2-small-en	33M	0.370	0.351	0.268	0.387
msmarco-distilbert-base-v4	66M	0.333	0.385	0.281	0.307
multilingual-e5-base	278M	0.333	0.514	0.438	0.347
multilingual-e5-large	560M	0.444	**0.558**	**0.449**	0.373
mxbai-embed-large-v1	335M	0.333	0.429	0.304	0.307
nomic-embed-text-v1	n/a	0.370	0.424	0.337	0.320
nomic-embed-text-v1.5	137M	0.370	0.424	0.338	0.307
paraphrase-mpnet-base-v2	n/a	0.444	0.364	0.317	0.387
paraphrase-multilingual-mpnet-base-v2	278M	0.296	0.349	0.239	0.253
qwen3-embedding-0.6b	596M	0.370	0.496	0.388	0.280
qwen3-embedding-4b	4B	0.407	0.512	0.407	0.280
qwen3-embedding-8b	7.6B	0.250	0.557	0.432	0.264
snowflake-arctic-embed-l-v2	568M	0.370	0.424	0.315	0.267
Lexical baseline					
tfidf	n/a	**0.444**	**0.424**	**0.262**	**0.227**

Table 13: **Length-bucket Success@5 on AuthBench (full results).** Authorship attribution is evaluated with Success@5 (S@5), Recall@5 (R@5), and nDCG@5 (higher is better). Authorship verification is evaluated with equal error rate (EER; lower is better). **Bold** = best, underlined = second best within each model group. **Size** denotes parameter count when publicly available.

Model	Size	short	medium	long	extra_long
LLMs (instruction-tuned)					
deepseek-coder-6.7b-instruct	6.7B	0.077	0.071	0.091	0.157
deepseek-llm-7b-chat	7B	0.077	0.079	0.109	0.200
llama3-8b-instruct	8B	0.077	**0.062**	0.086	0.101
llama3.1-8b-instruct	8B	0.093	0.064	0.083	**0.100**
qwen2.5-3b-instruct	3.1B	**0.065**	0.064	**0.074**	0.105
qwen2.5-7b-instruct	7.6B	0.069	0.072	0.076	**0.100**
qwen3-4b-instruct	4B	0.102	0.080	0.088	0.114
LLMs (base)					
deepseek-llm-7b-base	7B	**0.051**	0.077	0.095	0.157
llama3-8b	8B	0.103	**0.067**	0.087	**0.101**
llama3.1-8b	8B	0.103	**0.067**	0.087	0.106
qwen2.5-3b	3.1B	0.072	**0.067**	**0.077**	0.104
qwen3-4b	4B	0.103	0.082	0.085	0.113
Embedding models (instruction-tuned)					
e5-mistral-7b-instruct	7.1B	0.119	**0.083**	0.098	0.114
gte-qwen2-7b-instruct	7.6B	**0.112**	0.084	**0.087**	**0.099**
sfr-embedding-mistral	7.1B	0.116	**0.083**	0.098	0.114
Embedding models					
all-minilm-l12-v2	33M	0.103	0.117	0.185	0.228
all-minilm-l6-v2	23M	0.141	0.107	0.176	0.253
all-mpnet-base-v2	109M	0.103	0.109	0.165	0.228
all-roberta-large-v1	355M	0.205	0.102	0.126	0.215
allenai-specter	110M	0.081	0.126	0.164	0.256
bert-base-uncased	110M	0.116	**0.090**	**0.102**	0.152
bge-base-en-v1.5	109M	0.124	0.118	0.201	0.316
bge-base-zh-v1.5	102M	0.100	0.103	0.152	0.200
bge-large-en-v1.5	335M	0.112	0.106	0.180	0.316
bge-large-zh-v1.5	326M	0.207	0.099	0.164	0.190
bge-m3	568M	0.231	0.215	0.284	0.367
bge-small-en-v1.5	33M	0.154	0.151	0.215	0.304
distiluse-base-multilingual-cased-v2	135M	0.133	0.169	0.230	0.228
e5-base-v2	109M	0.128	0.131	0.191	0.253
e5-large-v2	335M	0.153	0.146	0.195	0.253
e5-small-v2	33M	0.139	0.163	0.237	0.278
facebook-contriever	109M	0.112	0.103	0.134	0.254
facebook-contriever-msmarco	109M	0.128	0.125	0.193	0.266
gte-base	n/a	0.128	0.107	0.175	0.278
gte-large	335M	0.104	0.094	0.149	0.266
gte-large-en-v1.5	434M	0.103	0.092	0.132	0.203
jina-embeddings-v2-base-en	137M	0.205	0.282	0.317	0.389
jina-embeddings-v2-small-en	33M	0.099	0.104	0.113	**0.141**
msmarco-distilbert-base-v4	66M	0.179	0.186	0.265	0.276
multilingual-e5-base	278M	0.154	0.134	0.165	0.177
multilingual-e5-large	560M	0.103	0.131	0.152	0.190
mxbai-embed-large-v1	335M	0.115	0.108	0.195	0.316
nomic-embed-text-v1	n/a	0.119	0.099	0.176	0.262
nomic-embed-text-v1.5	137M	0.114	0.092	0.154	0.249
paraphrase-mpnet-base-v2	n/a	0.114	0.112	0.165	0.266
paraphrase-multilingual-mpnet-base-v2	278M	0.179	0.226	0.273	0.304
qwen3-embedding-0.6b	596M	0.077	0.100	**0.102**	0.176
qwen3-embedding-4b	4B	**0.053**	0.096	0.104	0.171
qwen3-embedding-8b	7.6B	0.100	0.120	0.118	0.176
snowflake-arctic-embed-l-v2	568M	0.103	0.169	0.212	0.266
Lexical baseline					
tfidf	n/a	**0.081**	**0.136**	**0.187**	**0.316**

Table 14: **Length-bucket EER on AuthBench (full results).** Authorship attribution is evaluated with Success@5 (S@5), Recall@5 (R@5), and nDCG@5 (higher is better). Authorship verification is evaluated with equal error rate (EER; lower is better). **Bold** = best, underlined = second best within each model group. **Size** denotes parameter count when publicly available.

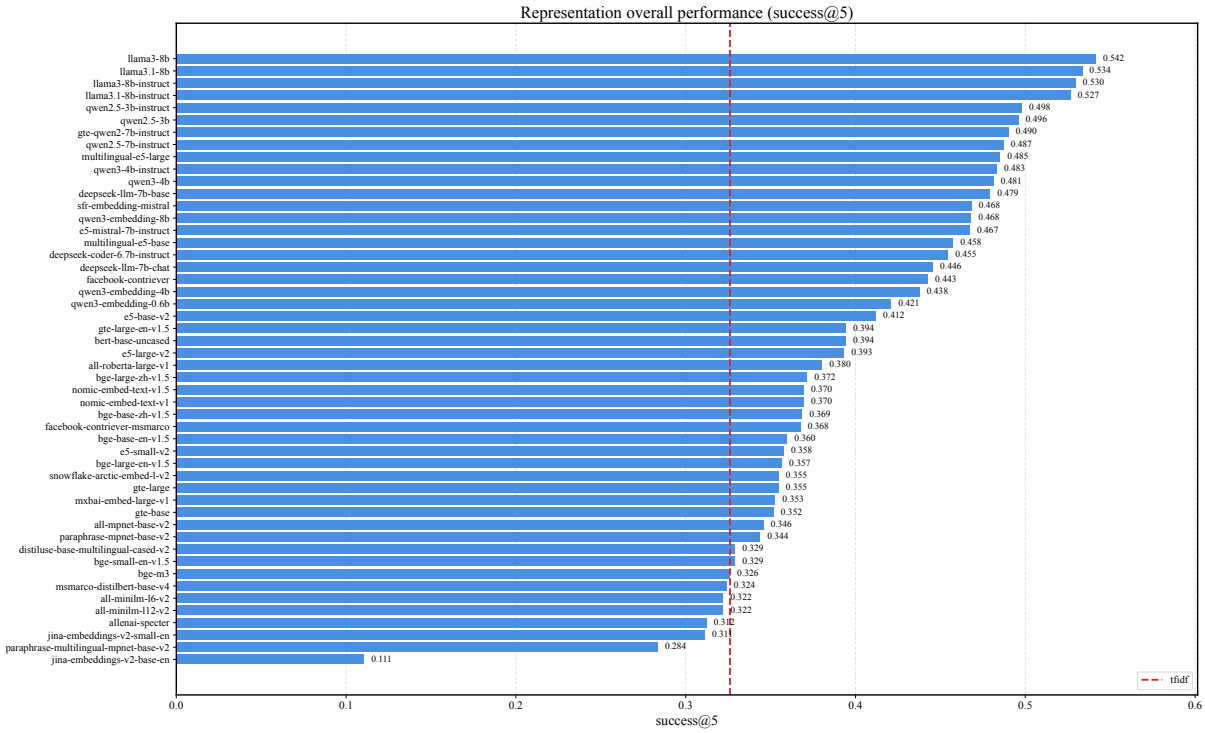

Figure 5: Macro-average of Sucess@5

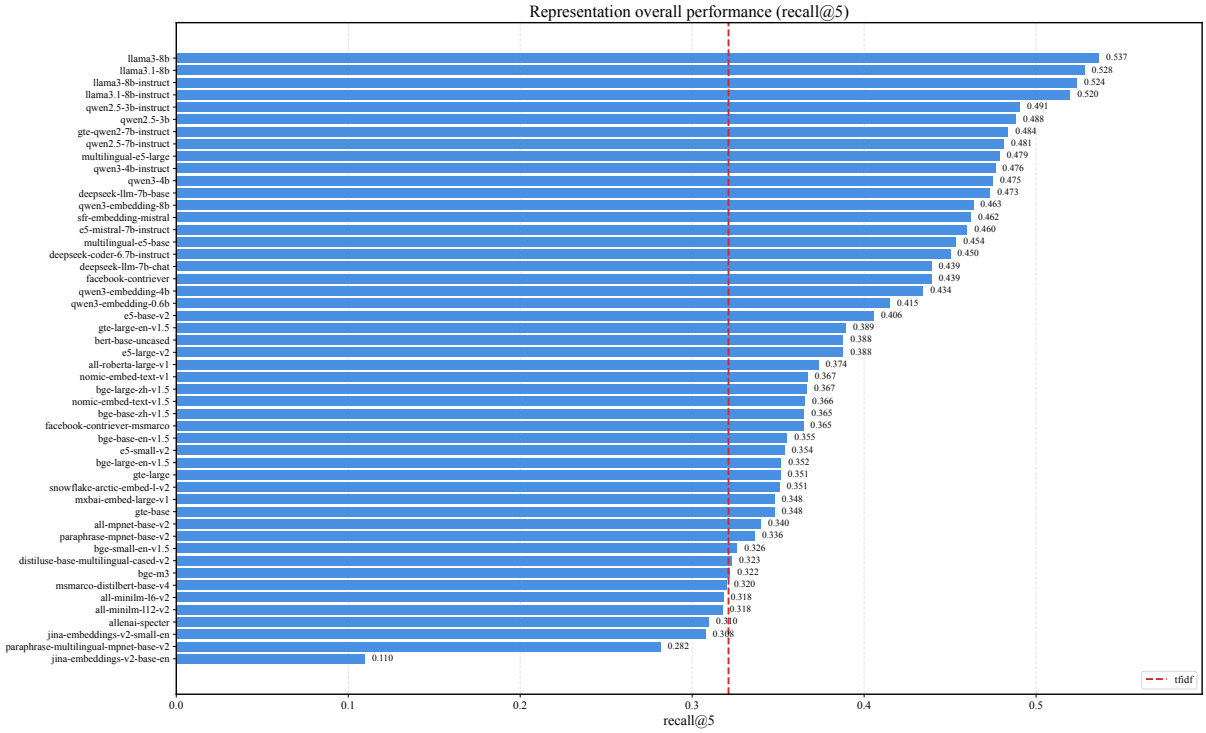

Figure 6: Macro-average of Recall@5

tracking, and explicit documentation of prohibited uses.

F.2 Statistics for Data

We report dataset scale, composition, and splits as follows.

- **Dataset size and composition:** Table 2 reports total documents/authors, language dis-

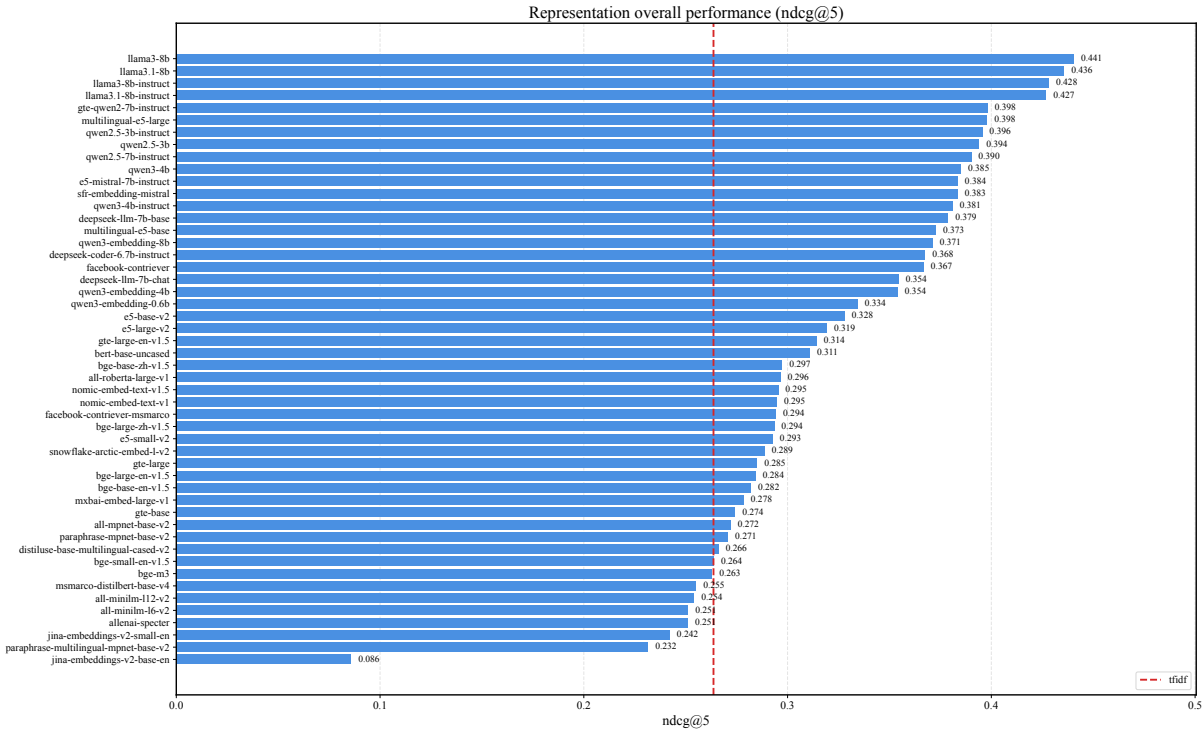

Figure 7: Macro-average of nDCG@5

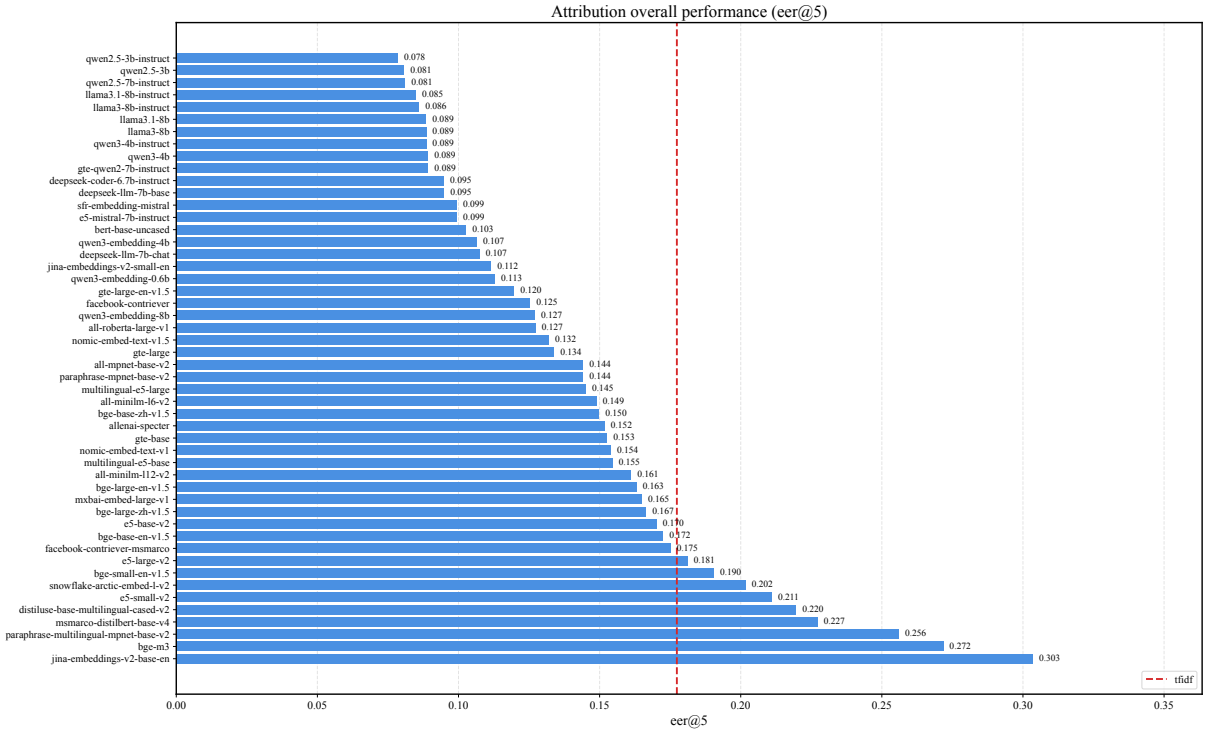

Figure 8: Macro-average of EER

tribution, and length-bucket counts, and Section 3 summarizes the benchmark contents and metadata.

- **Schema:** Section 3 describes the unified record schema and required metadata fields.
- **Splits and leakage reduction:** Section 3 and Appendix A.7 describe split construction and deduplication-aware assignment to reduce leakage; the full results tables are in Appendix B.

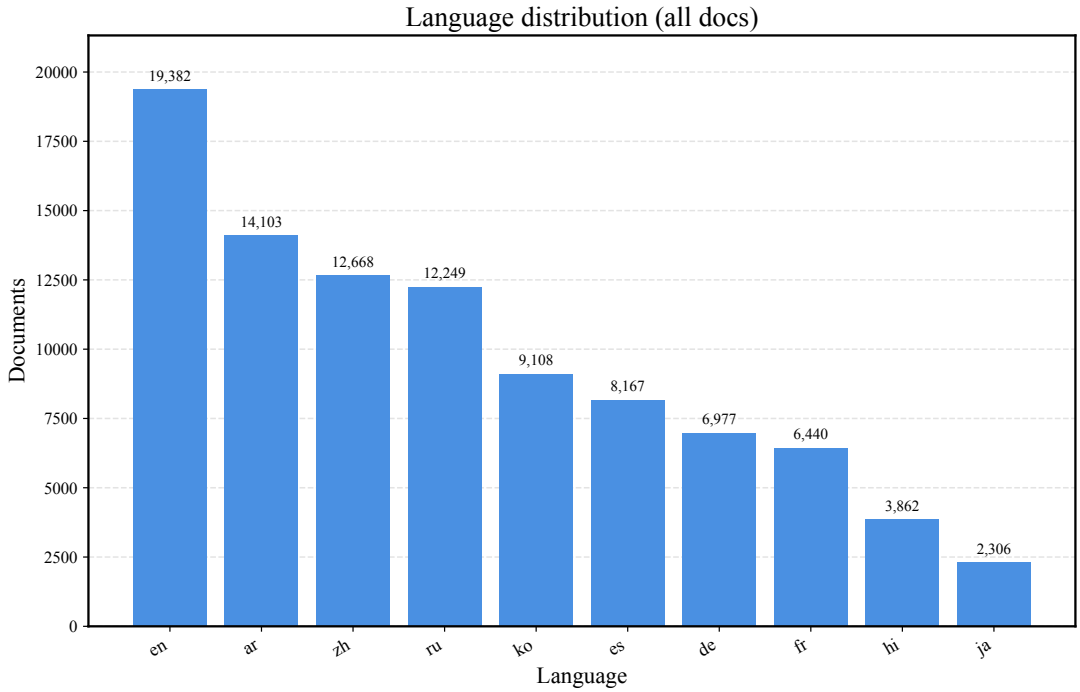

Figure 9: **Language distribution of AuthBench.** Bar heights show the number of documents per language.

- **Raw sources:** Appendix E lists all raw corpora and references.

F.3 Experimental Setup and Hyperparameters

Experimental setup. Our evaluation protocol is specified in Section A.8, including within-split candidate pools for retrieval and labeled query–candidate pairs for verification.

Hyperparameters. We evaluate off-the-shelf models without task-specific fine-tuning to ensure comparability (Section A.8). Accordingly, there is no hyperparameter search for model training in the reported baselines. All evaluation-side parameters (e.g., K for Success@K/Recall@K/nDCG@K; candidate pool construction; negative sampling options, if enabled) are fixed and implemented in the released evaluation toolkit (Section A.8).

Please make the table in a more standard and coherent format in my research paper. Fix the ”1” issue with other notation represenation

F.4 Descriptive Statistics

We report (i) aggregate metrics and (ii) mandatory stratified breakdowns by language, genre, and length (Section B). Because these baselines are evaluated deterministically (no stochastic training in this paper) and computed on fixed splits, we primarily report point estimates rather than confidence intervals. To support transparency about variability and slice composition, the evaluation toolkit additionally outputs per-slice sample sizes and diagnostics (e.g., number of queries per slice, candidate pool statistics). If future versions include stochastic training or sampling, we will add multiple-seed runs and uncertainty estimates (e.g., bootstrap confidence intervals across queries) here.

F.5 Topic-Leakage Controls and Diagnostics

A common concern in authorship evaluation is that systems exploit topical overlap or lexical duplication rather than author-specific signals. To make such shortcut behavior explicit, we include a sparse TF-IDF cosine baseline and stress-test it by modifying candidate pool construction while keeping evaluation strictly within split. Concretely, in addition to the default within-split candidate pool, we construct a topic-selected variant where candidates are restricted to match a coarse topic label (when available from source metadata or derived topic tags). Table 17 shows two complementary effects: (i) TF-IDF improves under topic selection (higher S@5), indicating that lexical matching can benefit from increased topical alignment; yet (ii) verification degrades (higher EER), consistent with lexical similarity being a poor proxy for calibrated authorship decisions. Importantly,

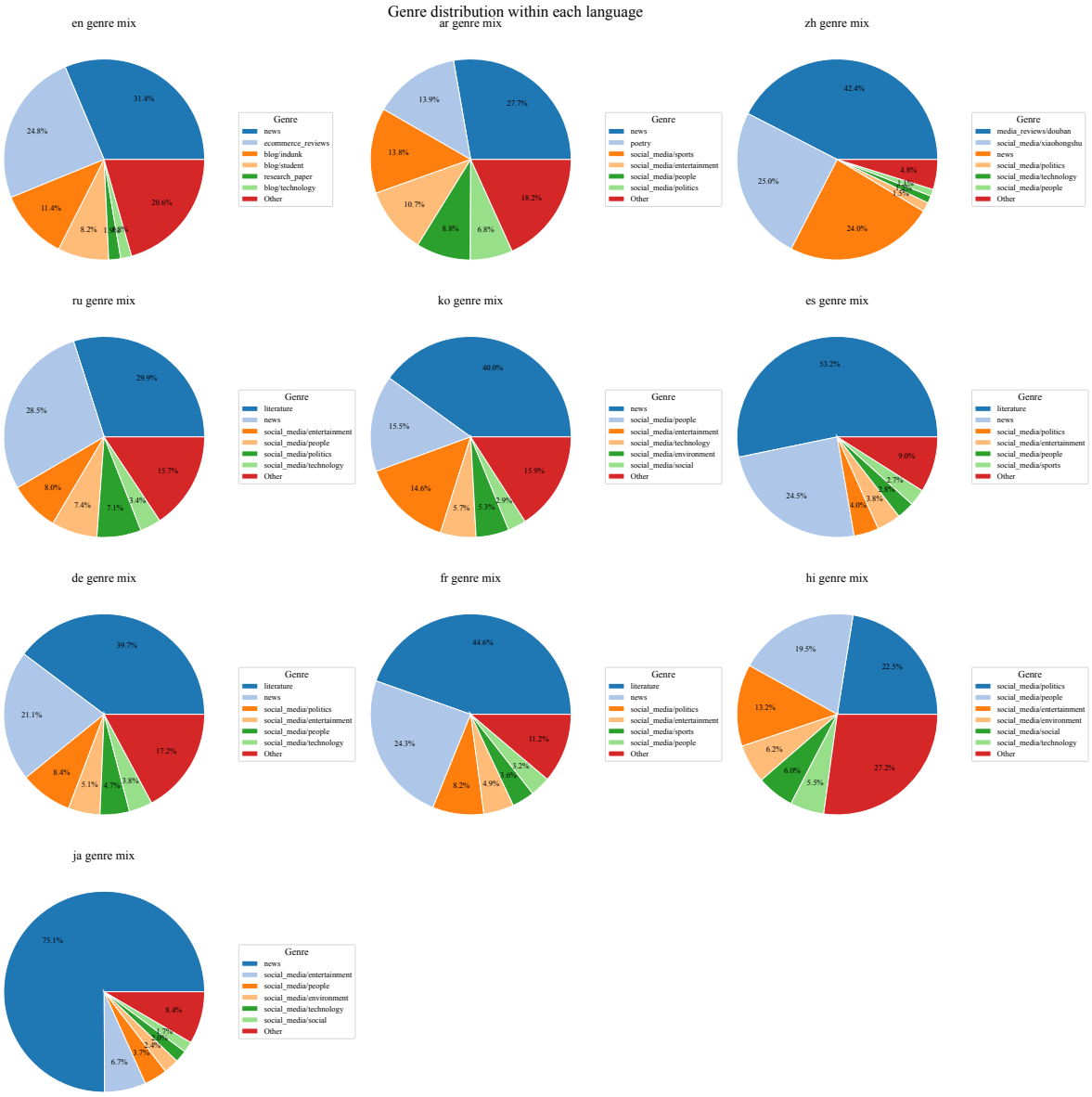

Figure 10: **Genre distribution per language.** Each subplot is one language.

TF-IDF remains substantially below strong neural baselines overall, suggesting that AuthBench is not dominated by trivial lexical overlap and remains robust and challenging under a diagnostic that would otherwise amplify topic shortcuts.

F.6 Instructions Given to Participants

This work does not involve human subjects experiments, crowdworker annotation, or participant-provided responses. Therefore, there are no participant instructions to report.

Human-in-the-loop quality checks. Although we do not use participants, we include a limited human-in-the-loop review for dataset quality and correctness as part of the construction pipeline. Reviewers inspected samples of normalized instances and slice distributions to (i) validate field mappings (language/genre/source), (ii) spot-check filtering behavior (e.g., obvious boilerplate, malformed text), and (iii) verify that anonymization and deduplication behaved as intended. This review is a research quality assurance step rather than a human-subject study and does not involve collecting new personal data.

F.7 Data Consent

AuthBench is derived from publicly available datasets released by their respective providers under documented licenses/terms (Appendix E). We did not collect new data directly from individuals,

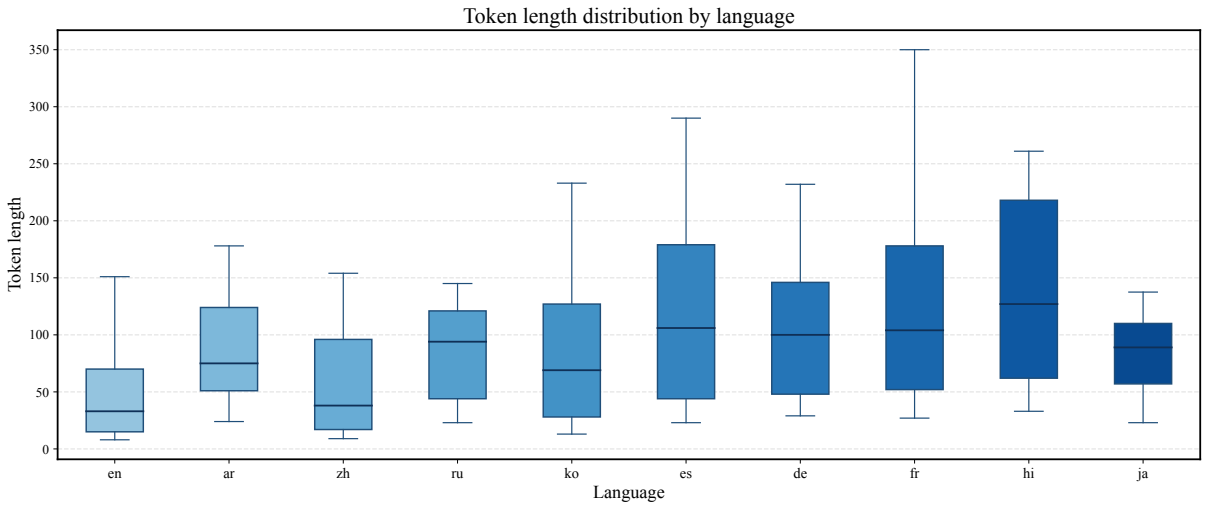

Figure 11: **Token-length distribution per language.** Each box plot summarizes token-length quantiles for one language using a box summary (box: p25–p75; whiskers: p10–p90; median marked).

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 Papers (metadata)	en	Research papers	1.7M+	Yes (author list)	(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)

Table 15: **Raw corpora used to construct AuthBench.** "Scale" reflects the approximate size of the raw source as documented by the source provider and may exceed the final benchmark after filtering, deduplication, and per-author constraints.

Table 16: **Licensing / terms and release mode by source.** "Release mode" indicates whether we redistribute normalized text (Tier A) or provide manifest-only reconstruction (Tier B). URLs are provided for reproducibility and access-date auditing.

Source	License / terms (summary)	Release mode	Notes / pointer
Exorde	MIT (per dataset card)	Tier A	See provider page for license.
Babel Briefings	CC BY-NC-SA 4.0 (per dataset card)	Tier A	Redistribution requires attribution; non-commercial and share-alike constraints apply.
Amazon Reviews Multi (MARC)	Research-only, non-commercial; no republishing	Tier B	Users must obtain the corpus from the original distributor; we provide manifests and reconstruction scripts only.
Blog Authorship Corpus	Non-commercial research use (as documented by provider)	Tier B	We release anonymized IDs and splits; reconstruction requires obtaining the source.
arXiv (metadata)	Metadata CC0 (incl. abstracts); full-text licenses vary	Tier B	We treat title/abstract as metadata; we do not redistribute PDFs; reconstruction via metadata access.
Xiaohongshu /Weibo	See hosting/provider terms	Tier B	If redistribution terms are unclear, default to manifest-only release.
Douban Reviews	See hosting/provider terms	Tier B	Same as above.
Hindi Discourse Stories	Provider-restricted; contact required for some uses	Tier B	Manifest-only unless explicit redistribution permission is documented.
Spanish/French/Russian/ German PD corpora	Public domain (per provider)	Tier A	Public-domain texts redistributed as normalized content.
Arabic Classical Poetry	Open data license as documented by provider	Tier A/B	Choose Tier A only if the documented license explicitly permits redistribution.

TF-IDF setting	S@5 ↑	EER ↓
Default candidate pool	0.4095	0.1802
Topic-selected candidate pool	0.4969	0.2361

Table 17: **Topic-leakage diagnostic with a lexical baseline.** We evaluate sparse TF-IDF under two within-split candidate-pool constructions. Even with a topic-selected pool, TF-IDF remains a comparatively weak baseline, supporting that AuthBench is not easily solved by surface lexical overlap alone.

and we did not contact data subjects. Where consent mechanisms are relevant, we rely on the original dataset providers' consent and terms for collection and redistribution. We further reduce privacy risk by anonymizing author identifiers (Section 3; Section 3.1) and applying conservative safety filtering (Appendix A.3).

F.8 Ethics Review Board Approval

We did not conduct new human-subject data collection or interventions. The benchmark is constructed from existing public datasets with established distribution terms (Appendix E). Accordingly, we did not seek separate IRB/ethics board approval for data collection. If your institution requires a formal determination, you may optionally obtain an exemption letter; otherwise, we treat this as not applicable to new data collection.

F.9 Information About Use of AI Assistants

Option A (if applicable): AI assistants used. We used AI assistants to support manuscript drafting and editing (e.g., improving clarity, restructuring paragraphs, and formatting LaTeX tables/figures), and to help implement non-novel engineering components of the evaluation pipeline. All methodological claims, experimental design decisions, dataset construction specifications, and reported results were designed, verified, and validated by the authors. AI assistants did not determine the benchmark's core scientific contributions or substitute for author verification of correctness.