LLM Summarization Attribution Report

Yujin Chen
University of Hawaii at Mānoa
yujin31@hawaii.edu

Abstract

Note: There was a typo in figures where it wrote "Granite-3.1" when it was supposed to be "Granite-3.3."

I study multi-class authorship attribution for large language models in a constrained summarization setting. Summaries are generated from the same source documents using five models (Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Qwen-2.5-7B-Instruct, GLM-4-9B-Chat, and Granite-3.3-8B-Instruct) across four domains: News, Research, Reddit, and Email. Because all summaries describe the same content, attribution must rely on differences in wording and structure rather than topic.

I train a RoBERTa-large classifier to predict the source model from each summary. The model achieves 88.27% accuracy on the test set. Performance varies by model, with Llama-3.1 being easiest to identify and Mistral-v0.3 and GLM-4 being the most confusable. Analysis of confusion patterns and word frequencies shows that many errors are driven by shared high-frequency and domain-specific words. These results indicate that model-specific fingerprints remain detectable under strong semantic constraints, but are weakened by overlapping domain vocabulary.

GitHub:

https://github.com/yujin-chen/LLM-Summarization-Attribution

Hugging Face:

https://huggingface.co/datasets/yujin31/model_generated_summaries

1 Introduction

Generative AI models have developed quickly in recent years. Initially its small set of closed commericial API's has expanded into a large ecosystem of open-weight and enterprise models. As these systems are deployed more widely, concerns around safety, misuse, and accountability have become increasingly important. One open problem is to determining which model generated a given piece of text.

My project focuses on the problem of multi-class model authorship attribution. Given a set of candidate language models and a generated summary, the goal is to predict which model produced the text. Formally, given a set of candidate models $\mathcal{M} = \{M_1, ..., M_5\}$ and a generated summary S, I aim to train a classifier f_θ such that $f_\theta(S) \rightarrow \hat{y}$, where $\hat{y}$ is the correct source model.

Attribution is especially difficult for modern language models because many share similar architectures, training data, and alignment techniques. In addition, there is often no metadata or watermark attached to generated text, making verification challenging (Pasquini et al., 2025).

In this project, I focus on authorship attribution in context of summarization. Summarization constrains the model to the content of the source document, reducing semantic variation across outputs. All summaries describe the same facts, meaning the classifier cannot rely on topic or meaning to make predictions. Instead, it must learn subtle differences in wording, structure, and token usage.

With these setup I come up with the following research question:

- Do stylistic fingerprints remain visible when semantic content is fixed?
- Can a classifier distinguish between models of similar size (7B–9B parameters) that differ mainly in training focus, such as chat-oriented versus enterprise-oriented models?

1.1 Problem

Although there has been significant work on detecting AI-generated text, attributing text to a specific model remains difficult. There are three main challenges.

- There is no reliable way to verify the origin of generated text after it is produced. This makes it hard to distinguish whether content came from a safety-aligned enterprise model or a more general-purpose chat model.
- In summarization tasks the semantic content is fixed by the source document. Because all summaries describe the same information, the classifier cannot rely on factual differences. Any successful attribution must depend on subtle statistical patterns in wording, syntax, and token choice.
- Modern language models often converge toward similar behavior due to shared architectures and training data. This raises the question of whether meaningful model-specific signals still exist, and if so, whether they can be detected reliably.

2 Related Work

Research in finding the authors of text has moved from studying human writing style (Stamatatos, 2009) to identifying the source of Large Language Models (LLMs). Recent methods mostly use active fingerprinting, where models are tested with specific prompts to reveal unique behaviors. For example, Pasquini et al. (Pasquini et al., 2025) introduces *LLMmap*, which finds models using carefully designed questions, while Pei et al. (Pei et al., 2025) look at thinking styles and behaviors like being overly agreeable (sycophancy).

However, these methods typically depend on direct model access or open-ended generation. My research addresses this limitation by focusing on passive attribution within constrained summarization. Unlike active fingerprinting, I assume a 'black-box' scenario where only the final output is observable. By analyzing summaries where the semantic content is fixed, I strictly test whether unique stylometric signatures persist even when the model's creative freedom is restricted.

3 Preliminaries: Models

This study considers five large language models with similar parameter scales (7B–9B) but distinct training objectives, data mixtures, and tokenization strategies. All models are evaluated in a zero-shot summarization setting using identical prompts.

Llama-3.1-8B-Instruct, instruction-tuned model released by Meta. It's for general-purpose conversational use, trained to adhere closely to user instructions and safety guidelines(Meta AI, 2024).

Mistral-7B-Instruct-v0.3, instruction-tuned model developed by Mistral AI. It utilizes a transformer architecture optimized for efficiency and possesses a larger vocabulary size than its predecessors. This model represents a compact, highly capable open-weight system with behavioral patterns often compared to Llama, providing a test case for distinguishing architecturally similar models(Mistral AI, 2024).

Qwen-2.5-7B-Instruct, multilingual instruction-tuned model released by Alibaba Cloud. It is trained on a massive, diverse corpus covering nearly 30 languages and code. Crucially, Qwen employs a distinct tokenizer with a larger vocabulary than Llama-style models, which significantly influences word segmentation and surface-level token probability distributions(Qwen Team, 2024).

GLM-4-9B-Chat, general-purpose language model developed by Zhipu AI. Optimized for English-Chinese bilingual tasks, it utilizes a unique training setup and tokenization scheme that differs from standard Llama-based architectures(Zhipu AI, 2024).

Granite-3.3-8B-Instruct, enterprise-focused language model released by IBM. Unlike the chat-centric models, Granite is trained on a curated mixture of finance, legal, and code data, prioritizing reliability and transparency over creative conversational flair(IBM Research, 2025).

Together, these models provide a controlled but diverse testbed for studying authorship attribution across models that are similar in size but differ in training data, alignment goals, and intended use.

4 Methodology

4.1 Data

I construct a multi-domain summarization dataset covering four domains: News articles(Hugging Face Datasets, 2025), Academic research papers(Hugging Face Datasets, 2018), Reddit posts(Hugging Face Datasets, b), and Emails(Hugging Face Datasets, a). Source text are retrieved from different dataset from Hugging Face. For each source document, I generate five summaries, one from each model: Llama-3.1, Mistral-v0.3, Qwen-2.5, GLM-4, and Granite-3.3.

To reduce surface-level stylistic variation, all models are prompted using constraint-heavy in-

structions. These prompts specify summary length, formatting, and tone, discouraging obvious cues such as bullet points or informal phrasing. This setup is designed to normalize outputs and force the classifier to rely on finer-grained stylistic signals.

To prevent data leakage, summaries derived from the same source document never appear across different dataset splits. The source data are first splits into train, validation, test sets. Then generate summaries by splits. The final dataset consists of 100,000 training summaries, 20,000 validation summaries, and 20,000 test summaries. For train splits, its 5,000 summaries per model per domain. Validation and test splits they are both 1,000 summaries per model per domain.

4.2 Attribution Model

I view model attribution as a five-label supervised classification problem (one for each LLM). I fine-tune **RoBERTa-large** (Liu et al., 2019) as a sequence classifier to predict the source model from each generated summary. The model reads the raw text and learns patterns in subword usage and phrasing directly from data, without handcrafted features.

I implement training using the Hugging Face `transformers` library. I use the default RoBERTa-large tokenizer with dynamic padding. I use accuracy as the primary evaluation metric on the test set, and monitor macro-averaged precision, recall, and F1 score during validation.

The main training settings are as follows:

- Base model: RoBERTa-large for sequence classification.
- Number of classes: 5.
- Optimizer: AdamW.
- Learning rate: 3.85×10^{-5}.
- Weight decay: 0.2.
- Warmup: 0.1 warmup ratio.
- Epochs: 5 (with early stopping patience = 2).
- Batch size: 32 per device for training, 64 per device for evaluation.
- Model selection: best checkpoint selected by validation accuracy.

4.3 Further analysis

Beyond the standard accuracy metrics, I perform several analyses to understand model behavior. This include:

1. Confusion Analysis: Identifying specific model-pair confusions
2. Word Frequency Analysis: Comparing token frequencies in error cases to those in the full dataset, to identify words that commonly appear in misclassifications.

5 Results

5.1 Overall Performance

The RoBERTa-large classifier achieved an overall attribution accuracy of 88.27% across the test set. This result confirms that distinct stylometric fingerprints persist across 7B–9B parameter models, even when semantic content is strictly controlled by the summarization task.

Models vary in their performance. With a recall of 0.98, which indicates that nearly all of its summaries are correctly classified, Llama-3.1 is the most easily recognized model. Granite-3.3 and Qwen-2.5 show moderate recall at 0.89 and 0.87, respectively. Mistral-v0.3 and GLM-4 are the most difficult to distinguish, with recalls of 0.84 and 0.83. These findings suggests that attribution difficulty varies depending on how different each model's generation patterns are.

5.2 Confusion Matrix and Misclassification Patterns

The model exhibits different levels of identifiability across the five architectures, as seen in Figure 1.

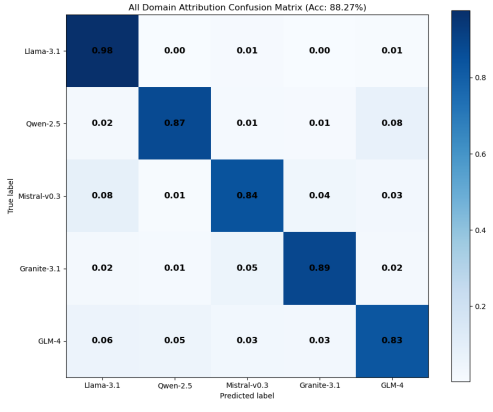

Figure 1: Confusion Matrix of the 5-way attribution task.

With a nearly perfect true positive rate (recall) of 0.98, Llama-3.1 stands out as being extremely separable. Only a small percentage of its summaries are incorrectly classified, namely as GLM-4 (27 cases) or Mistral-v0.3 (44 cases). This implies that consistent style cues seen in Llama summaries are rarely mistaken for those found in other models.

On the other hand, Mistral-v0.3 is sometimes mistaken for Llama-3.1. Mistral summaries are anticipated as Llama in 334 occasions, while the contrary is true in just 44 cases. This significant disparity demonstrates that while Llama is still more unique, several Mistral summaries fit into Llama's learned feature space.

Another major error pattern involves Qwen-2.5 and GLM-4. Specifically, 306 Qwen summaries are misclassified as GLM-4, accounting for roughly 8% of all Qwen samples. The stronger Qwen-to-GLM direction suggests that GLM's broader lexical patterns frequently absorb Qwen's stylistic cues, even though GLM-4 is also confused with Qwen (197 cases).

Granite-3.3 shows a different failure mode. Although its overall recall is high, it is most often confused with Mistral-v0.3, with 217 Granite summaries misclassified as Mistral. This confusion is stronger than Granite's confusion with Llama (96 cases) or GLM (80 cases), suggesting a closer stylistic overlap between Granite and Mistral.

GLM-4 has the weakest separability overall. Its errors are spread across all other models, with notable confusion toward Llama-3.1 (238 cases) and Qwen-2.5 (197 cases). This broad error distribution suggests that GLM-4 shares common lexical patterns with multiple models rather than closely matching a single one.

These patterns are clearly visible in the error heatmap shown in Figure 2, which highlights the strong directionality of the largest misclassification flows. May refer to Appendix B for domain specific matrices.

5.3 Lexical Evidence: Signature vs. Deceptive Tokens

To better understand why these errors occur, I examine word frequency statistics and the vocabulary of misclassified summaries.

Across all domains, the most frequent words used by each model show substantial overlap. For example, Mistral-v0.3 uses the word user 2,712 times and energy 1,848 times, while Granite-3.3 uses energy 2,186 times and user 1,878 times.

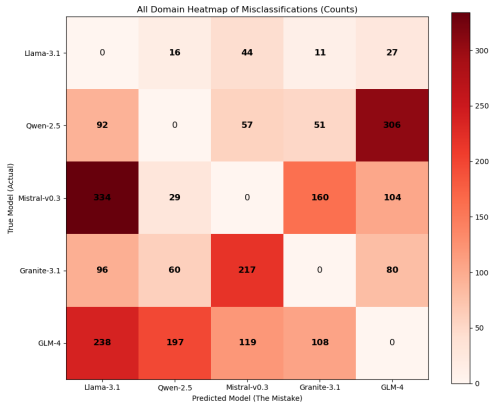

Figure 2: Error Heatmap (Misclassification Counts). This figure highlights the directionality of errors.

GLM-4 frequently uses including (1,030) and despite (993), while Qwen-2.5 strongly favors poster (1,274) and like (1,048). These numbers indicate that many high-frequency words are shared across models, limiting their usefulness for attribution.

The deceptive-word analysis shows that many of these same high-frequency words also dominate misclassified examples. In the all-domain error table, energy appears as a top error-trigger word for four out of five models, with counts ranging from 132 (Qwen-2.5) to 212 (Mistral-v0.3). Other common error-trigger words include including, new, power, and state. Because these words appear frequently regardless of the generating model, they tend to confuse the classifier rather than help it distinguish authorship.

5.4 Cross-Domain Dynamics

Error patterns become clearer when broken down by domain.

In the Email domain, errors are dominated by Enron-related vocabulary (see Appendix A.1). Words such as enron, company, california, million, and market appear hundreds of times across all models. For instance, enron appears 783 times for Llama, 732 times for Mistral, 554 times for Qwen, 497 times for GLM, and 946 times for Granite in error-related statistics. This shared domain vocabulary overwhelms stylistic differences and leads to frequent cross-model confusion.

Moving on to News domain, the errors are driven by reporting conventions (see Appendix A.2). Words like year, old, year old, incident, police, and numeric placeholders such as 000 appear repeatedly across all models. For example, year appears 460 times for Llama and 589 times for Granite in

error cases. These terms reflect the structure of news writing rather than model-specific style.

Then Research domain, the errors are associated with shared scientific terms such as model, study, field, energy, and quantum(see Appendix A.3). These words appear thousands of times in the full test set (e.g., model appears 1,203 times for Llama and 1,554 times for Granite), making it difficult for the classifier to rely on vocabulary alone.

Finally, Reddit domain shows a different pattern. Platform-related words such as user, poster, seeking, and advice dominate both the frequency tables and the error tables (see Appendix A.4). For example, user appears 2,649 times for Mistral and 1,828 times for Granite. These framing words provide some attribution signal, but when multiple models adopt similar forum-style phrasing, they also become a major source of confusion.

Llama-3.1	Mistral-v0.3	Qwen-2.5	GLM-4	Granite-3.3
energy (62)	energy (212)	poster (158)	user (199)	user (261)
state (56)	user (200)	power (140)	including (172)	million (249)
distribution (52)	new (188)	new (135)	poster (166)	energy (206)
power (44)	power (188)	energy (132)	new (163)	new (195)
time (42)	including (164)	enron (122)	despite (161)	power (182)
gas (42)	enron (157)	including (115)	energy (156)	enron (165)
8a0yaaab... (42)	million (143)	2001 (110)	state (117)	round (132)
multiplicity dist. (38)	potential (132)	state (101)	power (115)	including (123)
multiplicity (38)	model (128)	despite (97)	million (111)	despite (118)
new (35)	time (126)	like (84)	year (109)	gas (115)

Table 1: Top 10 'Deceptive Words' per Model (All-Domain). High-frequency triggers causing misclassification.

Llama-3.1	Mistral-v0.3	Qwen-2.5	GLM-4	Granite-3.3
including (1780)	user (2712)	poster (1274)	energy (1125)	energy (2186)
energy (1636)	energy (1848)	like (1048)	study (1030)	user (1878)
new (1436)	new (1463)	energy (1022)	including (1030)	model (1649)
model (1292)	model (1378)	new (953)	new (994)	including (1475)
power (1166)	power (1243)	including (861)	despite (993)	study (1466)
state (1161)	study (1143)	despite (820)	poster (871)	new (1455)
study (1113)	state (1138)	power (728)	potential (768)	despite (1303)
user (993)	time (1085)	using (652)	like (766)	potential (1283)
poster (941)	potential (1058)	potential (645)	power (717)	power (1268)
using (937)	using (1048)	high (636)	user (698)	paper (1245)

Table 2: Top 10 Most Frequent Words per Model (All Sample). Shows the 'native vocabulary' of each model.

Tables 1 and 2 summarize the all-domain lexical patterns that drive attribution behavior. The baseline frequency table shows that several high-frequency words are shared across models (e.g., *energy*, *new*, *including*, *user*), while the deceptive-word table demonstrates that many of these same tokens also appear frequently in misclassified examples. This overlap explains why domain-heavy summaries often lead to systematic attribution errors.

5.5 Fingerprint Robustness and Sequence Length

Figure 3 shows how attribution accuracy varies with summary length. Very short summaries (under 100 tokens) have the lowest accuracy at 0.84, indicating that limited text provides fewer stylistic cues. Accuracy improves steadily as summaries become longer, reaching 0.91 for 200–300 tokens and peaking at 0.95 for summaries between 400 and 500 tokens.

For summaries longer than 500 tokens, accuracy slightly decreases to 0.90. This suggests that while more text generally helps attribution, very long summaries may introduce additional generic content that weakens model-specific signals.

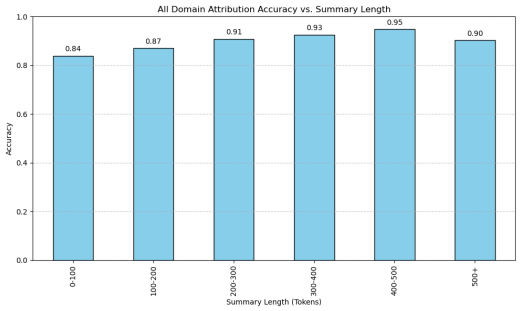

Figure 3: Attribution Accuracy vs. Summary Length.

6 Discussion

RQ1: Do stylistic fingerprints remain visible when semantic content is fixed?

Yes, even when the models were forced to write about the exact same topic, their unique styles remained visible. My classifier achieved 88.27% accuracy, successfully telling these similar-sized models apart despite the strict constraints of the summarization task. Surprisingly, this worked even on very short texts (under 100 words). This suggests that LLM fingerprints are "holographic"—meaning a model's identity is hidden in small details like word choice, rather than needing a long document to reveal itself.

RQ2: Can a classifier distinguish between models of similar size (7B–9B) that differ mainly in training focus?

Yes, but the classifier used different clues depending on the model's focus. For the enterprise-focused Granite, the detection was partly based on a shortcut. Granite tended to repeat specific keywords (like "Enron") much more than the others, allowing the classifier to spot it based on content rather than just style.

In contrast, the chat-focused models were identified by their actual writing patterns, though they sometimes blurred together. For example, Mistral often copied Llama's formal tone, while Qwen and GLM used the exact same internet slang. In these cases, the models shares traits that made them harder to distinguish.

7 Next Steps

My current results show that attribution works well overall, but many errors are linked to shared high-frequency words and domain terms. A reasonable next step is to test whether the classifier is learning true model style or mainly using these lexical cues.

I will run masking experiments where I remove or replace the most common domain words and the top error-trigger words, then retrain and re-evaluate. If accuracy drops sharply, it would suggest the model relies heavily on dataset artifacts. If performance stays high, it would support the claim that model fingerprints are more stable and not purely lexical.

I will also test cross-domain generalization by training on three domains and evaluating on a held-out domain. This directly measures whether fingerprints transfer across different writing styles and vocabularies.

References

Hugging Face Datasets. a. Enron email dataset. https://huggingface.co/datasets/LLM-PBE/enron-email.

Hugging Face Datasets. b. Reddit dataset (reddit_ds_479243). https://huggingface.co/datasets/zkpbeats/reddit_ds_479243.

Hugging Face Datasets. 2018. arxiv summarization dataset. https://huggingface.co/datasets/ccdv/arxiv-summarization.

Hugging Face Datasets. 2025. CNN/DailyMail news summarization dataset. https://huggingface.co/datasets/abisee/cnn_dailymail.

IBM Research. 2025. Granite-3.3-8b-instruct model card. https://huggingface.co/ibm-granite/granite-3.3-8b-instruct.

Yinhan Liu and 1 others. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Meta AI. 2024. Llama-3.1-8b-instruct model card. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Mistral AI. 2024. Mistral-7b-instruct-v0.3 model card. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Dario Pasquini, Evgenios M. Kornaropoulos, and Giuseppe Ateniese. 2025. Llmmap: Fingerprinting for large language models. *Preprint*, arXiv:2407.15847.

Zehua Pei, Hui-Ling Zhen, Ying Zhang, Zhiyuan Yang, Xing Li, Xianzhi Yu, Mingxuan Yuan, and Bei Yu. 2025. Behavioral fingerprinting of large language models. *Preprint*, arXiv:2509.04504.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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

Zhipu AI. 2024. Glm-4-9b-chat model card. https://huggingface.co/zai-org/glm-4-9b-chat.

A Frequency Table

This appendix provides detailed frequency tables for the top 10 lexical features associated with each model. The data is stratified by domain to highlight specific vocabulary artifacts (e.g., "Enron" in Email vs. "poster" in Reddit) that influence attribution performance.

A.1 Email Domain

Llama	Mistral	Qwen	GLM	Granite
state (42)	power (228)	enron (107)	energy (108)	million (327)
enron (35)	enron (207)	power (106)	power (99)	energy (215)
energy (30)	energy (183)	energy (101)	state (86)	enron (203)
power (27)	million (142)	2001 (94)	california (80)	power (175)
gas (26)	company (121)	california (75)	enron (80)	round (174)
said (26)	california (110)	state (72)	new (55)	new (167)
new (22)	market (107)	san (69)	market (50)	gas (118)
private (17)	state (106)	francisco (63)	million (49)	com (111)
com (16)	new (94)	san francisco (63)	including (48)	market (110)
companies (16)	gas (92)	chronicle (62)	company (45)	california (103)

Table 3: Top 10 'Deceptive Words' per Model (Email Domain Error).

Llama	Mistral	Qwen	GLM	Granite
power (889)	power (939)	power (593)	energy (614)	energy (1022)
enron (783)	energy (787)	energy (584)	power (563)	enron (946)
energy (769)	enron (732)	enron (554)	enron (497)	power (946)
company (625)	email (617)	include (387)	new (372)	new (703)
state (536)	million (592)	new (383)	market (330)	million (688)
new (488)	new (581)	email (354)	million (345)	market (494)
million (485)	gas (427)	million (345)	including (298)	gas (491)
including (479)	california (423)	including (336)	million (270)	california (484)
california (430)	com (419)	california (306)	gas (246)	email (462)
market (421)	market (414)	market (267)	email (241)	2001 (445)

Table 4: Top 10 'Deceptive Words' per Model (Entire Email Data).

A.2 News Domain

Llama	Mistral	Qwen	GLM	Granite
doolittle (4)	year (93)	despite (37)	despite (85)	despite (75)
kroeger (4)	despite (86)	new (37)	including (68)	year (70)
mission (4)	old (64)	including (23)	year (67)	old (56)
cole (3)	including (63)	000 (20)	000 (48)	year old (49)
attack (3)	league (58)	team (19)	old (48)	new (46)
medal (3)	team (58)	public (18)	police (47)	team (43)
museum (3)	police (54)	league (17)	new (45)	including (41)
raid (3)	family (53)	year (17)	year old (44)	incident (39)
bombing (2)	year old (53)	million (16)	team (40)	expressed (34)
saturday (2)	new (51)	face (15)	million (32)	match (34)

Table 5: Top 10 'Deceptive Words' per Model (News Domain Error).

Llama	Mistral	Qwen	GLM	Granite
year (460)	year (416)	despite (387)	despite (580)	despite (589)
team (423)	despite (325)	year (258)	year (375)	year (463)
including (396)	new (307)	including (212)	including (282)	including (363)
old (348)	including (289)	new (207)	old (274)	old (353)
year old (299)	league (263)	like (206)	year old (240)	year old (315)
new (294)	old (259)	old (201)	team (239)	new (287)
incident (254)	expressed (254)	team (182)	new (226)	team (282)
league (243)	team (236)	year old (175)	expressed (204)	league (271)
despite (227)	000 (234)	league (154)	league (195)	000 (260)
000 (213)	years (219)	000 (152)	000 (176)	incident (242)

Table 6: Top 10 'Deceptive Words' per Model (Entire News Data).

A.3 Research Domain

Llama	Mistral	Qwen	GLM	Granite
model (27)	study (116)	xcite (47)	model (54)	magnetic (43)
time (27)	model (100)	systems (39)	study (53)	study (32)
energy (21)	results (84)	model (35)	ray (47)	model (29)
study (21)	field (66)	asymptotics (31)	observations (46)	neutrino (28)
symmetry (19)	energy (65)	star (31)	systems (46)	continuum (27)
dust (18)	high (65)	quantum (30)	field (44)	energy (25)
field (18)	quantum (63)	data (28)	energy (38)	fields (25)
nuclear (18)	ray (63)	generalization (28)	high (37)	models (25)
clusters (17)	paper (58)	nonintegrable systems (27)	models (37)	field (24)
matter (17)	non (56)	spin (27)	using (34)	using (24)

Table 7: Top 10 'Deceptive Words' per Model (Research Domain Error).

Llama	Mistral	Qwen	GLM	Granite
model (1203)	model (1307)	model (535)	study (950)	model (1554)
study (1025)	study (1033)	models (446)	model (625)	study (1367)
energy (820)	energy (1006)	energy (407)	quantum (541)	paper (1215)
results (765)	field (813)	high (404)	energy (472)	energy (1105)
quantum (757)	results (766)	quantum (401)	models (464)	using (874)
field (746)	high (724)	using (383)	systems (462)	field (861)
including (648)	mass (714)	like (381)	high (434)	models (852)
mass (614)	using (708)	systems (372)	research (416)	mass (834)
using (608)	quantum (682)	study (350)	analysis (399)	quantum (816)
used (606)	paper (668)	non (327)	field (377)	research (814)

Table 8: Top 10 'Deceptive Words' per Model (Entire Research Data).

A.4 Reddit Domain

Llama	Mistral	Qwen	GLM	Granite
user (22)	user (167)	poster (141)	user (140)	user (377)
poster (20)	game (37)	despite (28)	poster (129)	poster (99)
new (9)	mother (20)	user (28)	game (49)	seeking (67)
sie (9)	ve (19)	game (26)	ve (45)	game (63)
art (8)	like (18)	seeking (25)	seeking (40)	despite (60)
seeking (7)	new (18)	like (24)	using (38)	ve (56)
und (7)	poster (18)	advice (22)	including (36)	like (48)
ist (6)	potential (18)	including (22)	new (32)	new (48)
felt (5)	games (17)	ve (22)	despite (29)	advice (44)
moonpup (5)	including (17)	new (19)	like (29)	potential (40)

Table 9: Top 10 'Deceptive Words' per Model (Reddit Domain Error).

Llama	Mistral	Qwen	GLM	Granite
user (949)	user (2649)	poster (1274)	poster (870)	user (1828)
poster (941)	seeking (416)	like (283)	user (669)	poster (491)
ve (392)	game (284)	despite (254)	seeking (335)	seeking (355)
game (315)	advice (270)	advice (218)	ve (297)	despite (332)
looking (312)	new (270)	new (217)	despite (260)	ve (295)
new (302)	potential (244)	seeking (198)	game (223)	like (289)
like (284)	user seeking (199)	ve (186)	new (222)	new (252)
seeking (260)	seeking advice (191)	game (181)	like (219)	game (241)
including (257)	like (188)	using (157)	advice (204)	advice (236)
experience (199)	using (186)	issues (145)	including (176)	potential (194)

Table 10: Top 10 'Deceptive Words' per Model (Reddit Domain Entire Test Data).

B Confusion Matrices and Error Heatmaps

This appendix provides domain-specific confusion matrices and error heatmaps for the five-way authorship attribution task. Each domain includes one normalized confusion matrix and one misclassification heatmap.

The confusion matrices show normalized prediction behavior, while the heatmaps show raw misclassification counts. Together, they highlight both accuracy and directionality of errors.

B.1 Email Domain

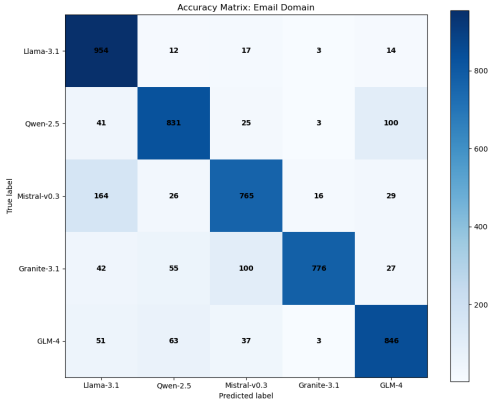

Figure 4: Normalized confusion matrix for the Email domain.

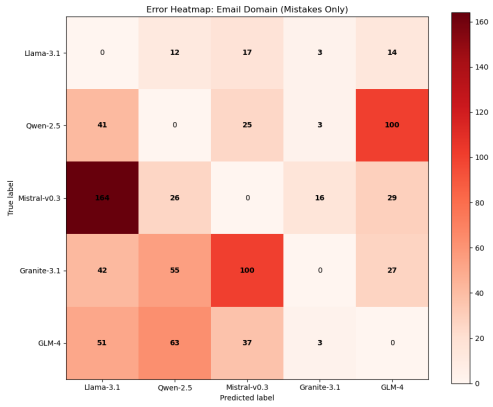

Figure 5: Error heatmap (misclassification counts) for the Email domain.

B.2 News Domain

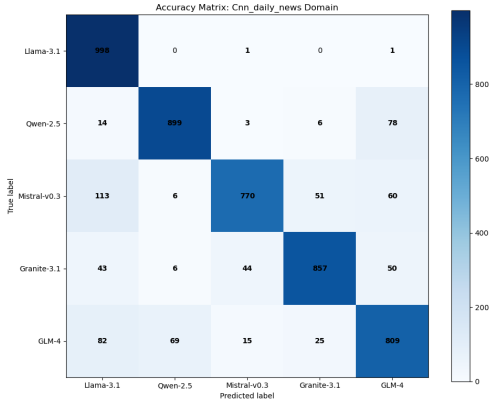

Figure 6: Normalized confusion matrix for the News domain.

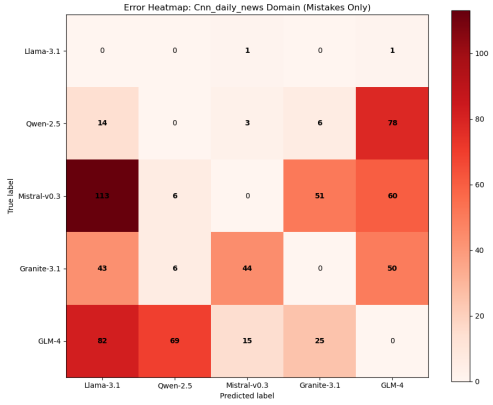

Figure 7: Error heatmap (misclassification counts) for the News domain.

B.3 Research Domain

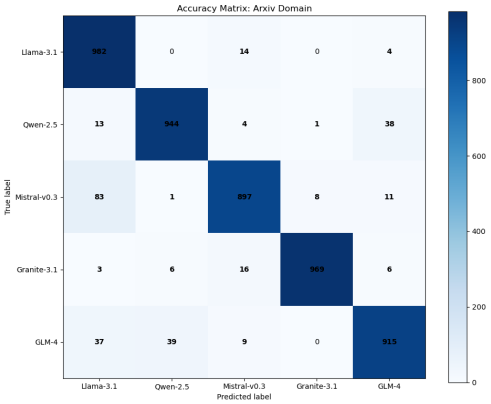

Figure 8: Normalized confusion matrix for the Research domain.

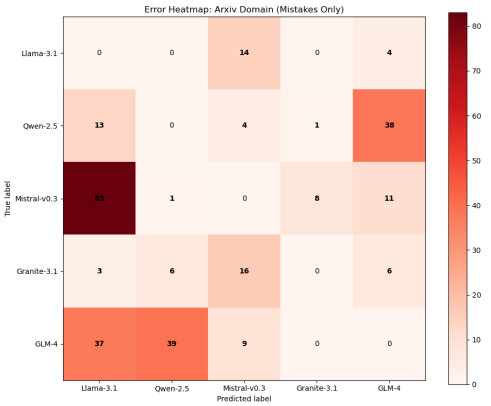

Figure 9: Error heatmap (misclassification counts) for the Research domain.

B.4 Reddit Domain

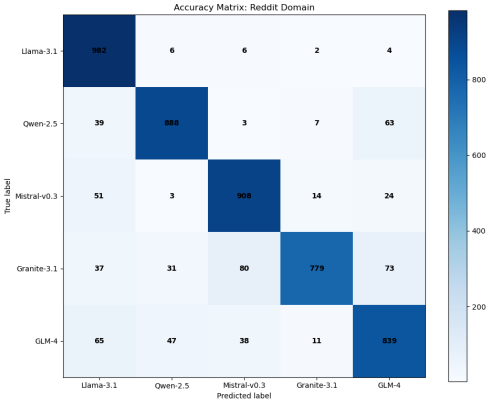

Figure 10: Normalized confusion matrix for the Reddit domain.

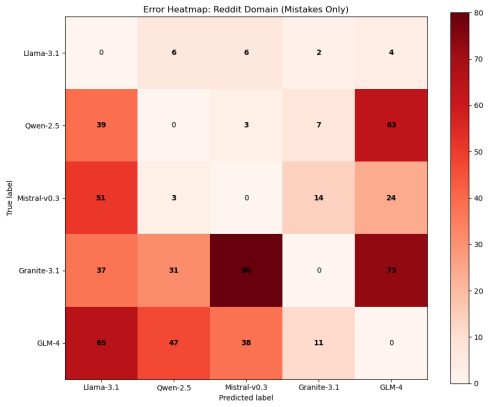

Figure 11: Error heatmap (misclassification counts) for the Reddit domain.