Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Retrieval-augmented large language models for bank complaint prioritization: accuracy, bias transfer and fairness-aware retrieval

Fahad Alsheref

fabdulwahed@kku.edu.sa

King Khalid University

Research Article

Keywords: retrieval-augmented generation, large language models, complaint management, algorithmic fairness, banking analytics, decision support

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11200080/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Retrieval-augmented large language models for bank complaint prioritization: accuracy, bias transfer and fairness-aware retrieval

Fahad Kamal Alsheref

*Department of Business Informatics, College of Business, King Khalid University, Abha, Saudi Arabia.

Corresponding author(s). E-mail(s): fabdulwahed@kku.edu.sa;

Abstract

Banks must decide which customer complaints to escalate, and large language models (LLMs) now offer to automate this decision. Zero-shot LLMs rank complaints that end in monetary relief close to chance, because the outcome depends on institutional history that the complaint text does not contain. This study tests whether retrieval-augmented generation (RAG), which supplies similar past complaints and their outcomes, closes that gap, and whether it transfers historical disparities between consumer groups. Using 309,348 complaints from 2021–2024 as retrieval memory and complaints from 2025–2026 as queries, drawn from the Consumer Financial Protection Bureau archive, we evaluate Qwen3-8B and Claude Haiku 4.5 under six context conditions against three non-LLM baselines. On a pilot set, similar cases with their outcomes raise ROC AUC from 0.61 to 0.76–0.77 for both models; similar cases without outcomes, or random cases with outcomes, give no gain. On the 3,000-complaint test set, Qwen3-8B rises from 0.56 to 0.81, matching supervised models (0.80–0.82) but not the historical relief rate used directly as a score (0.84). Institutional memory also widens the recall gap between servicemembers and other consumers, from +0.6 percentage points without context to **−4.1** with full context, approaching the **−6.0** to **−8.1** points of the base rate and supervised models, because servicemembers' meritorious complaints retrieve neighbours with lower historical relief rates. For complaint prioritization, what the model is shown matters more than which model is used, and retrieval memory requires a fairness audit before deployment.

Keywords: retrieval-augmented generation, large language models, complaint management, algorithmic fairness, banking analytics, decision support

1 Introduction

Complaint prioritization determines which customers receive faster and more senior attention when they report a problem with a bank. Supervisors expect banks to classify, escalate and resolve complaints within fixed deadlines; Saudi Arabia's central bank, for example, requires resolution within five working days together with defined escalation procedures and root-cause analysis [1]. Large language models (LLMs) promise to automate this work because they read free-text narratives without task-specific training [2, 3].

A companion study found that this promise holds for only half of complaint triage [4]. LLMs assigned complaints to the correct product with the highest accuracy of all models tested, but

ranked complaints that ended in monetary relief close to chance (ROC AUC 0.51–0.62), while supervised models trained on past outcomes reached 0.80–0.82. The study attributed the gap to institutional knowledge: whether a complaint ends in relief depends on how companies have resolved similar cases, and the complaint text does not reveal it. The supervised models that captured this knowledge also inherited a historical disparity and flagged fewer relief complaints from servicemembers than from other consumers, echoing documented inequities in complaint handling [5].

Retrieval-augmented generation (RAG) offers a direct test of that explanation. RAG supplies a model with retrieved documents at inference time [6], and banks increasingly deploy LLMs this way because it adds institutional knowledge without retraining. If institutional history is what zero-shot LLMs lack, retrieving similar past complaints and their outcomes should close the priority gap. The same mechanism raises a fairness concern. Recent work shows that RAG can undermine the fairness of aligned LLMs even when the retrieved corpus appears unbiased [7, 8], and that the composition of in-context examples shifts LLM fairness [9, 10]. When the retrieved outcomes themselves encode a historical disparity, RAG may transfer that disparity directly into LLM decisions.

This study examines both effects with complaints about retail bank products from the archive of the US Consumer Financial Protection Bureau (CFPB). It asks four questions. RQ1: does retrieval of similar past complaints and their outcomes raise LLM priority performance to the level of supervised models? RQ2: does retrieval transfer the historical servicemember disparity into LLM prioritization? RQ3: which part of the retrieved context carries the signal? RQ4: does the LLM add value beyond simple statistics computed from the same memory?

The study makes three contributions. First, it shows that retrieved cases with outcomes raise LLM priority performance by about 0.15 ROC AUC for both an open-source and a commercial model, and isolates the outcomes as the source of the gain. Second, it shows that an LLM with full context matches supervised models but does not outperform the historical relief rate computed from the same memory, which questions the need for an LLM in this step. Third, it traces how retrieval transfers disparity: relief complaints from servicemembers retrieve neighbours with lower historical relief rates, and the servicemember recall gap widens as more historical outcome information enters the prompt.

2 Related work

2.1 LLMs for complaint and outcome classification

LLMs classify consumer-finance complaints by product with high zero-shot accuracy [2], and consumers now use LLMs to write the complaints that institutions receive [11]. Outcome prediction has relied on supervised learning: monetary relief can be predicted from narrative text, topics and company identity with gradient-boosted trees [12]. Across business tasks, LLMs label text reliably but fail when the target depends on information outside the text; Megahed et al. [3] report that 14 LLMs classify financial-news sentiment accurately yet predict market movements at chance, and Luketina et al. [13] find that LLMs support tax decisions but struggle with implicit, context-specific knowledge. Retrieval over financial complaints has been studied for privacy-preserving classification [14], but not as a source of institutional memory for outcome prediction.

2.2 Fairness of retrieval and in-context learning

RAG changes the fairness properties of LLMs. Hu et al. [7] show that retrieval can undermine the fairness alignment of LLMs without any fine-tuning, even when the external corpus is censored, and Wu et al. [8] find fairness differences across RAG pipelines in open-domain question answering. Fairness-aware retrieval re-ranks documents to control the bias they pass to the generator [15]. For in-context learning, the proportion of minority-group examples in a prompt changes fairness outcomes [9], and demonstration selection can be optimized for fairness [10, 16]. These studies use benchmark bias datasets with demographic or synthetic labels. None examines retrieval of an organization's own historical decisions, where the retrieved labels carry a documented disparity.

2.3 Label bias

When a model predicts a proxy outcome, disparities in how the proxy was generated pass into its predictions [17]. Monetary relief is such a proxy for complaint merit: it records what companies granted, including any unequal treatment of consumer groups. Equality of opportunity, which requires equal true-positive rates across groups, is the natural criterion when a positive prediction confers a benefit [18]. Retrieval makes label bias immediate, because the LLM reads the historical labels in its prompt.

3 Data and study design

3.1 Data

We use the CFPB Consumer Complaint Database narratives archive [19], which preserves complaint narratives published before the CFPB stopped publishing them on 14 August 2026 [20]. Following [4], we keep complaints received from 1 January 2021 that include a narrative, concern one of five retail bank product classes (deposit accounts, credit cards, mortgages, consumer loans and money transfers) and carry a final company response. The outcome is whether the company closed the complaint with monetary relief.

3.2 Retrieval memory and query sets

The retrieval memory contains all 309,348 complaints received between 2021 and 2024. All query complaints were received from 1 January 2025 to 14 August 2026, so every retrieved case precedes its query and no future information enters the prompt. We use three query sets: (i) a pilot set of 500 randomly drawn complaints (relief rate 9.2%), used for condition comparisons and for tuning decision thresholds; (ii) the tag-stratified test set of [4], with 1,000 untagged complaints, 1,000 tagged "Older American" and 1,000 tagged "Servicemember", reweighted to population shares; and (iii) 2,100 supplementary relief complaints (700 per group) from the same period, used to estimate recall by group. The three sets are disjoint.

3.3 Retrieval

We embed every narrative with the bge-base-en-v1.5 sentence encoder [21], truncated to 256 tokens and normalized, and retrieve the k most similar memory complaints by cosine similarity with an exact FAISS index [22]. The main analyses use $k = 8$. The mean cosine similarity between a query and its nearest memory complaint is 0.87, and 73.5% of retrieved cases share the query's product class.

3.4 Context conditions

All conditions share one instruction and one output schema, an integer from 0 to 100 estimating the likelihood of monetary relief, and differ only in the context placed before the new complaint (Table 1). Each context case shows the company, issue, the first 700 characters of the narrative and, where applicable, its outcome. The new complaint shows its company, issue and narrative. No condition reveals the consumer tag.

Table 1 Context conditions. All conditions include the new complaint's company, issue and narrative.

Condition	Context before the new complaint
C0	None (zero-shot)
C1	k random memory complaints with outcomes
C2	k most similar memory complaints with outcomes
C3	C2 plus the company–issue historical relief rate
C4	Company–issue historical relief rate only
C5	k most similar memory complaints without outcomes

The company–issue relief rate is computed from the memory with an m-estimate ($m = 20$) that shrinks sparse cells toward the company rate, and the company rate toward the overall rate.

3.5 Models and baselines

We evaluate Qwen3-8B [23], run locally with 4-bit quantization through Ollama at temperature 0 with reasoning disabled, and Claude Haiku 4.5, accessed through the Anthropic Message Batches API. Both models return JSON constrained to the output schema. We compare them with three

scorers that use no LLM: the share of relief among the k retrieved neighbours (kNN vote), the company–issue historical relief rate (base rate), and the TF-IDF logistic regression and fine-tuned DeBERTa priority models of [4].

3.6 Evaluation

We report ROC AUC and precision–recall AUC with 95% bootstrap confidence intervals (1,000 resamples on the pilot set; 2,000 for group gaps). For each scorer we set the flagging threshold that maximizes F1 on the pilot set and apply it unchanged to the test and supplementary sets. For RQ2 we report priority recall by consumer group on pooled relief complaints (test set and supplementary sample; 779 untagged, 843 older American and 785 servicemember) and the difference from the untagged group. To trace the mechanism we compare, by group, the share of relief among the neighbours retrieved for relief complaints.

3.7 Experimental setup and implementation

Table 2 lists every setting needed to reproduce the experiments. All local computation ran on one laptop with an Intel Core i9-14900HX processor, 32 GB of memory and an NVIDIA GeForce RTX 4070 Laptop GPU with 8 GB of memory. The Claude runs used the Anthropic Message Batches API, which processes requests asynchronously at half the standard price.

Table 3 reports prompt length, latency and cost by condition. Retrieval quadruples the prompt: a zero-shot prompt contains about 380 tokens for Qwen3-8B and 580 for Haiku 4.5, and a prompt with eight retrieved cases about 1,800 and 2,100. Embedding the memory took 7.6 minutes on the GPU. Exact search over the memory on the CPU takes about 30 ms for a single query and about 1 ms per query when queries are processed in batches. The Haiku 4.5 pilot (1,500 requests) consumed 2.36 million input and 16,500 output tokens and cost US$1.22. Every LLM response in the pilot returned a valid score.

Appendix A reproduces the instruction and the structure of each prompt.

3.8 Use of generative AI

Besides the LLMs under study, the author used Claude Opus 5.5 (Anthropic), through the Claude Code application, to write and debug the analysis code, to search for and verify bibliographic records, and to draft and edit the manuscript. The author designed the study, verified all code, results and citations, and takes full responsibility for the content.

4 Results

4.1 Retrieved outcomes close most of the priority gap (RQ1, RQ3)

Table 4 reports priority performance on the pilot set. Without context (C0), both LLMs rank relief complaints weakly (ROC AUC 0.61), reproducing the zero-shot result of [4]. Retrieving the eight most similar past complaints with their outcomes (C2) raises ROC AUC to 0.74 for Qwen3-8B and 0.73 for Haiku 4.5, and adding the historical relief rate (C3) raises it to 0.77 and 0.76. The two models, which differ in size and provider, follow the same pattern within 0.02.

The ablations locate the signal in the retrieved outcomes. Similar cases without their outcomes (C5) give no gain over zero-shot (0.564), and random cases with outcomes (C1) give none either (0.564). The gain therefore requires both relevance and outcomes: the model benefits from knowing how similar cases ended, and neither topical similarity nor generic outcome examples substitute for that.

4.2 The LLM matches supervised models but not the base rate (RQ4)

Table 5 reports priority performance on the population-weighted test set. Qwen3-8B rises from 0.561 without context (C0) to 0.728 with retrieved cases (C2) and 0.805 with retrieved cases and the base rate (C3). Under C3 it matches the supervised models of [4] (DeBERTa 0.801, TF-IDF 0.817) and exceeds the kNN vote over the same neighbours (0.745). The company–issue base rate used directly as a score, however, reaches 0.835, the highest ROC AUC of all scorers, without any LLM call or model training.

Table 2 Experimental settings.

Component	Setting
Data and splits	
Retrieval memory	309,348 complaints received 1 Jan 2021–31 Dec 2024
Pilot set	500 complaints received 1 Jan 2025–14 Aug 2026, random draw (seed 42), 46 relief cases
Test set	3,000 complaints, 1,000 per consumer group, population-reweighted
Supplementary relief set	2,100 relief complaints, 700 per group (seed 7)
Retrieval	
Encoder	BAAI/bge-base-en-v1.5 (768 dimensions), 16-bit inference, batch size 128
Input length	First 256 tokens of each narrative; embeddings L2-normalized
Index	FAISS `IndexFlatIP` (exact inner-product search, equal to cosine similarity)
Neighbours	$k = 8$ in all main conditions; 32 stored per query
Context truncation	700 characters per retrieved case; 2,500 characters for the new complaint
Base rate	Company–issue relief rate with m-estimate smoothing, $m = 20$ (12,750 cells)
Qwen3-8B	
Model	`qwen3:8b`, 8.2B parameters, Q4_K_M quantization, Ollama 0.34.4
Decoding	Temperature 0, reasoning disabled, context window 8,192 tokens, JSON schema output; output capped at 40 tokens for the test runs
Claude Haiku 4.5	
Model	`claude-haiku-4-5` via Message Batches API
Decoding	Default sampling, `max_tokens` 50, JSON schema structured output
Evaluation	
Threshold	F1-maximizing value on the pilot set, grid 0.01–0.99 in steps of 0.01
Confidence intervals	Percentile bootstrap: 1,000 resamples (pilot AUC), 2,000 (group gaps); seed 42
Software	
Libraries	Python 3.12.10, PyTorch 2.11.0 (CUDA 12.8), sentence-transformers 6.1.0, FAISS 1.15.1, scikit-learn 1.9.1, pandas 3.0.6, Anthropic Python SDK 1.8.0

Table 3 Prompt length, latency and cost per condition on the pilot set. Latency is the median per complaint for Qwen3-8B on the local GPU; cost is the Batches API price for Haiku 4.5.

	Qwen3-8B (local)		Claude Haiku 4.5	
Condition	Prompt tokens	Seconds	Prompt tokens	US$ per 1,000
C0 zero-shot	379	0.39	580	0.32
C1 random cases	1,722	_[a]	–	–
C2 similar cases	1,790	0.92	2,095	1.08
C3 C2 + base rate	1,814	0.92	2,118	1.09
C4 base rate only	402	0.38	–	–
C5 similar, no outcomes	1,731	0.90	–	–
kNN vote, base rate	No LLM call; retrieval about 30 ms per query			

[a] C1 ran while another process shared the GPU, so its latency is not comparable.

The pilot results show why the LLM adds little once the memory is in the prompt. Given only the base rate (C4), Qwen3-8B reaches 0.691 on the pilot set, below the 0.801 that the same rate achieves as a score, so the model partly discards the signal it receives. Both LLMs also compress their estimates into 15–27 distinct values across 500 complaints, which limits their ability to rank.

4.3 Institutional memory transfers the servicemember disparity (RQ2)

The servicemember recall gap widens as more historical outcome information enters the prompt (Table 6, Figure 1). Without context (C0), Qwen3-8B flags servicemembers' relief complaints at the same rate as untagged consumers' (+0.6 percentage points). With retrieved cases (C2) the gap becomes −1.5 points, and with cases and the base rate (C3) −4.1 points. The base rate alone gives −6.0 points (95% CI −10.8 to −0.9), and

Table 4 Priority performance on the pilot set ($n = 500$, relief rate 9.2%). ROC AUC with 95% bootstrap confidence intervals. [a]PR AUC for Qwen3-8B, or for the scorer in rows without an LLM.

Scorer	Qwen3-8B	Claude Haiku 4.5	PR AUC[a]
C0 zero-shot	0.608 [0.52, 0.69]	0.613 [0.55, 0.67]	0.127
C1 random cases	0.564 [0.48, 0.64]	–	0.108
C2 similar cases	0.742 [0.67, 0.82]	0.730 [0.66, 0.80]	0.215
C3 C2 + base rate	0.768 [0.70, 0.83]	0.763 [0.69, 0.83]	0.220
C4 base rate only	0.691 [0.63, 0.76]	–	0.143
C5 similar, no outcomes	0.564 [0.48, 0.64]	–	0.101
Scorers without an LLM			
kNN vote ($k = 8$)	0.768 [0.69, 0.85]		0.265
Base rate	0.801 [0.73, 0.86]		0.288
TF-IDF (supervised)	0.824 [0.76, 0.88]		0.293

Table 5 Priority performance on the test set ($n = 3{,}000$, population-weighted). Thresholds were set on the pilot set.

Scorer	ROC AUC	PR AUC	Recall	Flag rate
Qwen3-8B C0 (zero-shot)	0.561	0.100	0.425	0.318
Qwen3-8B C2 (similar cases)	0.728	0.167	0.441	0.184
Qwen3-8B C3 (C2 + base rate)	0.805	0.201	0.473	0.164
kNN vote ($k = 8$), no LLM	0.745	0.200	0.350	0.153
Base rate, no LLM	**0.835**	0.274	0.390	0.114
TF-IDF (supervised)	0.817	**0.284**	0.604	0.212
DeBERTa (supervised)	0.801	0.238	**0.644**	0.234

the supervised models of [4] give −6.3 and −8.1 points. The individual LLM gaps have confidence intervals that include zero, but their ordering follows the amount of historical information in the context, and the fully informed scorers show statistically significant gaps. No scorer disadvantages older Americans.

The retrieval step explains the transfer. For complaints that did end in relief, the neighbours retrieved for servicemembers ended in relief 30.7% of the time, compared with 33.7% for untagged consumers and 36.1% for older Americans. Retrieval therefore presents a more pessimistic institutional memory for servicemembers' meritorious complaints, and a model that reads this memory receives the disparity as evidence.

5 Discussion

5.1 Context, not the model, drives prioritization

The results support the institutional-knowledge explanation of the LLM priority gap. Two unrelated LLMs gain the same 0.15 ROC AUC on the pilot set from retrieved outcomes and gain nothing from similarity without outcomes, and on the test set retrieval lifts Qwen3-8B from 0.56 to 0.81. The finding extends the observation of Megahed et al. [3] and Luketina et al. [13] that LLMs fail where the target depends on information outside the text: supplying that information fixes most of the failure. It also shows the limit of the LLM's contribution. The historical relief rate over the same memory ranks complaints better than the LLM that reads it, at no inference cost and with full transparency. For prioritization, a bank gains more from curating its resolution history than from choosing a larger model.

Table 6 Priority recall by consumer group on pooled relief complaints (test set and supplementary sample; 779 untagged, 843 older American, 785 servicemember) and gap from the untagged group in percentage points with 95% bootstrap confidence intervals. Thresholds were set on the pilot set.

	Recall		Gap vs untagged (pp)	
Scorer	Untagged	Servicem.	Servicemember	Older American
Qwen3-8B C0	0.436	0.442	+0.6 [−4.4, 5.7]	−2.2 [−6.9, 2.2]
Qwen3-8B C2	0.516	0.501	−1.5 [−6.5, 3.3]	+0.4 [−4.6, 5.1]
Qwen3-8B C3	0.531	0.490	−4.1 [−8.8, 0.9]	+2.3 [−2.4, 7.2]
kNN vote ($k = 8$)	0.499	0.459	−4.1 [−9.3, 0.9]	+4.3 [−0.4, 9.2]
Base rate	0.394	0.334	−6.0 [−10.8, −0.9]	+6.4 [1.4, 11.2]
TF-IDF (supervised)	0.641	0.577	−6.3 [−11.3, −1.6]	+3.1 [−1.5, 7.9]
DeBERTa (supervised)	0.646	0.564	−8.1 [−12.7, −3.4]	+5.7 [1.3, 10.3]

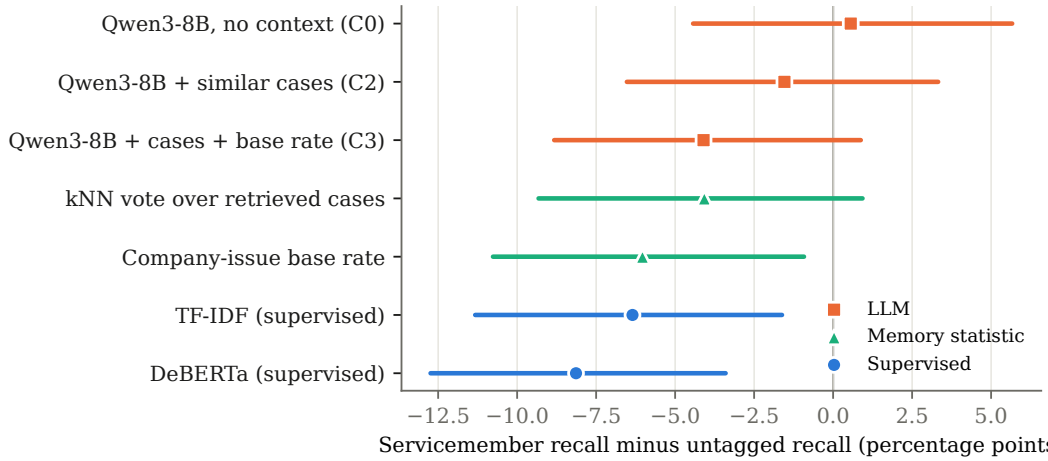


Fig. 1 The servicemember recall gap widens as the prompt carries more historical outcome information. Points show the gap from untagged consumers; bars show 95% bootstrap confidence intervals.

5.2 Retrieval memory needs a fairness audit

Retrieval does not create the servicemember disparity; it passes on the disparity recorded in past outcomes [5, 17]. The gradient in Figure 1 shows the trade-off directly: each step that adds institutional memory improves ranking and moves recall for servicemembers further below that of other consumers. The mechanism differs from the fairness failures reported for RAG on benchmark corpora [7, 8], where bias enters through retrieved text; here it enters through retrieved decisions. Two remedies follow from the literature: balancing retrieved cases so that outcome rates do not differ by group [9, 15], and group-specific thresholds that equalize recall [18]. Testing these remedies against the accuracy they cost is the natural next step.

5.3 Limitations

The pilot set is small (46 relief complaints), so pilot differences below about 0.05 ROC AUC are not resolved, and Claude Haiku 4.5 was evaluated on the pilot set only. The recall gaps for the LLM conditions are individually not statistically significant; the evidence for transfer rests on their ordering, on the significant gaps of the fully informed scorers and on the retrieval mechanism. We use one encoder, one value of k in the main analyses and one prompt design, and the test runs capped Qwen3-8B's output at 40 tokens, which the pilot runs did not. Monetary relief is a proxy for merit, consumer tags are optional and

self-reported, and the data cover English-language complaints to US institutions.

6 Conclusion

Retrieval-augmented generation closes the gap between LLMs and supervised models in bank complaint prioritization, and it does so through the outcomes of similar past cases. The same retrieval carries historical disparities into the model's decisions, and the historical relief rate computed from the memory ranks complaints better than the LLM that reads it. Banks that adopt RAG for complaint handling should treat their resolution history as the core asset, audit it for group disparities before it reaches any model, and consider whether an LLM is needed for the prioritization step at all.

Statements and Declarations

Acknowledgments. The author thanks the Deanship of Research and Graduate Studies at King Khalid University for funding this work under grant number GRP/1/47.

Funding. This work was funded by the Deanship of Research and Graduate Studies at King Khalid University under grant number GRP/1/47.

Competing interests. The author declares no competing interests.

Data availability. The complaint data are publicly available from the CFPB FOIA Electronic Reading Room [19]. Code and sample identifiers will be deposited in a public repository upon acceptance.

Use of generative AI. See Section 3.

Appendix A Prompts

Instruction (identical in all conditions).

> You help a US retail bank prioritize consumer complaints for escalation. Estimate how likely it is that the company will resolve the NEW complaint with monetary relief to the consumer (a refund, reimbursement, fee reversal or other payment). Return an integer from 0 to 100. Base your estimate on the evidence provided; relief is uncommon, so reserve high values for complaints that clearly resemble cases that ended in relief. Personal details are redacted as XXXX. Respond with the JSON object only.

Output schema.

A JSON object with one required integer field, `relief_likelihood`, and no other fields.

Context block (C1, C2, C3, C5).

A heading followed by k cases in the form:

```
Past complaint 1 (company: <company>;
issue: <issue>):
<first 700 characters of the narrative>
Outcome: CLOSED WITH MONETARY RELIEF or
closed without monetary relief
```

The heading reads "Past complaints most similar to the new complaint, with their outcomes" (C2, C3), "Randomly selected past complaints to this bank group, with outcomes" (C1) or "Past complaints most similar to the new complaint (outcomes not shown)" (C5, which omits the outcome line).

Base-rate line (C3, C4).

```
Historical base rate: <r>% of past
complaints about this company and issue
ended in monetary relief.
```

New complaint (all conditions, last).

```
NEW complaint (company: <company>; issue:
<issue>):
<first 2,500 characters of the narrative>
```

References

[1] Saudi Central Bank: Regulations for establishing customer care departments in banks. SAMA Rulebook, Circular No. 44069265, 21 March 2023. Accessed 25 September 2026 (2023). https://rulebook.sama.gov.sa/en/regulations-establishing-customer-care-departments-banks

[2] Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K.: Think before you classify: the rise of reasoning large language models for consumer complaint detection and classification. Electronics **14**(6), 1070 (2025) https://doi.org/10.3390/electronics14061070

[3] Megahed, F.M., Chen, Y.-J., Jones-Farmer, L.A., Lee, Y., Wang, J.B., Zwetsloot, I.M.: Reliable decision support with LLMs: a framework for evaluating consistency in binary text classification applications. Journal of Business Analytics, 1–19 (2026) https://doi.org/10.1080/2573234X.2026.2652281

[4] Alsheref, F.K.: Large language models for complaint triage in retail banking: accuracy, fairness and cost. Manuscript under review (2026)

[5] Sun, Y., He, D., Kejriwal, M.: Examining persistent inequities in financial complaint resolution for older Americans and veterans in the United States. PNAS Nexus **5**(4), 086 (2026) https://doi.org/10.1093/pnasnexus/pgag086

[6] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401 (2020). https://doi.org/10.48550/arXiv.2005.11401

[7] Hu, M., Wu, H., Zhu, R., Guan, Z., Guo, D., Qi, D., Li, S.: No free lunch: Retrieval-augmented generation undermines fairness in LLMs, even for vigilant users. In: Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 18145–18170 (2025). https://doi.org/10.18653/v1/2025.findings-emnlp.984

[8] Wu, X., Li, S., Wu, H.-T., Tao, Z.: Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems. arXiv preprint arXiv:2409.19804 (2024). https://doi.org/10.48550/arXiv.2409.19804

[9] Hu, J., Liu, W., Du, M.: Strategic Demonstration Selection for Improved Fairness in LLM In-Context Learning. arXiv preprint arXiv:2408.09757 (2024). https://doi.org/10.48550/arXiv.2408.09757

[10] Ma, H., Zhang, C., Bian, Y., Liu, L., et al.: Fairness-guided Few-shot Prompting for Large Language Models. arXiv preprint arXiv:2303.13217 (2023). https://doi.org/10.48550/arXiv.2303.13217

[11] Shin, M., Kim, J., Shin, J.: The adoption and efficacy of large language models in US consumer financial complaints. Nature Human Behaviour **10**(4), 669–680 (2026) https://doi.org/10.1038/s41562-026-02409-4

[12] Wang, Z., Zhu, S., Chen, X.: From complaint narratives to monetary relief: a hybrid machine learning framework for CFPB consumer complaints. arXiv preprint arXiv:2606.22664 (2026). https://doi.org/10.48550/arXiv.2606.22664

[13] Luketina, M., Benkel, A., Schuetz, C.G.: Using large language models for legal decision-making in Austrian value-added tax law: a comparative study. Journal of Business Analytics, 1–25 (2026) https://doi.org/10.1080/2573234X.2026.2720626

[14] Yadav, H., Srivastava, E., Kumar, S.: Federated learning for complaint classification: navigating non-IID textual data challenges and ensuring privacy and explainability. Journal of Business Analytics (2026) https://doi.org/10.1080/2573234X.2026.2708658

[15] Zhao, Y., Efthymiou, V., Nummenmaa, J., Stefanidis, K.: Fairness-Aware Retrieval Optimization for Retrieval-Augmented Generation. arXiv preprint arXiv:2605.15790 (2026). https://doi.org/10.48550/arXiv.2605.15790

[16] Bhaila, K., Van, M.-H., Edemacu, K., Zhao, C.: Fair In-Context Learning via Latent Concept Variables. arXiv preprint arXiv:2411.02671 (2024). https://doi.org/10.48550/arXiv.2411.02671

[17] Zanger-Tishler, M., Nyarko, J., Goel, S.: Risk scores, label bias, and everything but the kitchen sink. arXiv preprint arXiv:2305.12638 (2023). https://doi.org/10.48550/arXiv.2305.12638

[18] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. arXiv preprint arXiv:1610.02413 (2016). https://do

i.org/10.48550/arXiv.1610.02413

[19] Consumer Financial Protection Bureau: CFPB Consumer Complaint Database narratives archive. FOIA Electronic Reading Room. Page last modified 14 September 2026; accessed 24 September 2026 (2026). https://www.consumerfinance.gov/foia-requests/foia-electronic-reading-room/cfpb-consumer-complaint-database-narratives-archive/

[20] Consumer Financial Protection Bureau: The CFPB to cease discretionary publication of complaint narratives and visualizations. Press release, 14 August 2026. Accessed 24 September 2026 (2026). https://www.consumerfinance.gov/about-us/newsroom/the-cfpb-to-cease-discretionary-publication-of-complaint-narratives-and-visualizations/

[21] Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., Nie, J.-Y.: C-Pack: Packed Resources For General Chinese Embeddings. arXiv preprint arXiv:2309.07597 (2023). https://doi.org/10.48550/arXiv.2309.07597

[22] Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2021) https://doi.org/10.1109/TBDATA.2019.2921572

[23] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025). https://doi.org/10.48550/arXiv.2505.09388