On the Effectiveness and Robustness of Open-Weight LLMs for Danish Hate Speech Detection

OBS! This paper contains examples of hate speech that may be offensive to some readers.

Anonymous ACL submission

Abstract

We present a study of open-weight large language models for Danish hate speech detection. We evaluate four models (LLaMA, Mistral, Gemma, Qwen) across various prompting strategies, cross-prompt generalization, Danish orthographic effects, and robustness under balanced and imbalanced distributions. Under balanced evaluation, models achieve strong performance with minimal fine-tuning, with Gemma excelling on format-based prompts and Qwen showing consistent performance across theory-driven patterns. Surprisingly, zero-shot fine-tuning matches or exceeds few-shot performance while introducing fewer failures, suggesting that in-context examples may interfere with fine-tuned representations. However, imbalanced evaluation reflecting real-world distributions reveals substantial degradation, with Gemma maintaining the strongest performance under class skew. Theory-informed prompts grounded in linguistic frameworks prove more robust under class imbalance than simple format-based patterns. Cross-prompt generalization varies substantially by model, though format-constrained patterns consistently fail to transfer while semantically-grounded patterns show more robustness. ASCII transliteration of Danish characters (æ, ø, å → ae, oe, aa) significantly degrades performance, demonstrating that multilingual pre-training has established meaningful orthographic representations. These findings demonstrate that strong balanced performance does not guarantee real-world readiness and we recommend complementing balanced training with imbalanced evaluation to estimate deployment performance. Code and evaluation scripts are available.[1]

1 Introduction

The proliferation of online hate speech poses urgent societal challenges demanding effective automated detection. Danish represents a critical yet understudied case as a low-resource language with approximately six million speakers, it lacks the annotated corpora and specialized models available for English (Kirkedal et al., 2019). Danish statistics document increased hate crimes with approximately half occurring on digital platforms (Rigspolitiet, 2024), concerning given causal links between social media activity and real-world hate crimes (Müller and Schwarz, 2020). Open-weight LLMs offer a promising path forward due to their customizability, transparency, and sensitivity. However, their effectiveness for Danish hate speech detection remains unexamined. To our knowledge, this paper presents the first systematic investigation of LLM efficacy for Danish hate speech detection.

Research on LLM-based hate speech detection has focused on English revealing that effectiveness depends on many factors. Zhang et al. (2024b) show LLMs exhibit excessive sensitivity to linguistic cues and poor calibration; Melis et al. (2025) demonstrate that hate speech definitions substantially affect performance. Prompt design plays a critical role (White et al., 2023), with chain-of-thought strategies improving implicit hate detection (Nghiem and Daumé Iii, 2024). Class imbalance also remains central as hate speech constitutes only a small fraction of real-world content (Tonneau et al., 2025).

Research on Danish hate speech detection remains limited. Zhang et al. (2024a) introduce SnakModel, the first open Danish LLM, showing continued pre-training on native text outperforms multilingual baselines, but notes Danish lacks native instruction-tuning data. Müller-Eberstein et al. (2025) show native Danish data more than doubles LLM acceptance rates versus translated alternatives, highlighting cross-lingual transfer limitations. Contemporary LLMs also exhibit anglocentric biases (Pedersen et al., 2025; Müller-Eberstein et al., 2025). These findings motivate our investigations main findings:

[1] https://github.com/EAZUUZ/Danish-LLM-Hate

- Under balanced evaluation, open-weight LLMs achieve strong performance with minimal fine-tuning. Imbalanced distributions reveal substantial degradation.

- Theory-informed prompts demonstrate robustness under class imbalance. Cross-prompt generalization varies by model, but format-constrained patterns consistently fail to transfer, while semantically-grounded patterns show more robust generalization.

- ASCII transliteration of Danish characters (æ, ø, å → ae, oe, aa) degrades performance, indicating meaningful learned representations that should be preserved.

2 Methodology

Dataset. We construct a balanced subset of 850 samples (425 hate, 425 non-hate), partitioned into training (80%), validation (10%), and test (10%) splits using stratified sampling. All models are fine-tuned on the same balanced training set.

To assess robustness, we evaluate under two settings: (1) a *balanced* test set from the balanced subset, and (2) an *imbalanced* test set preserving the natural class distribution (13% hate). Additionally, we create ASCII-transliterated variants replacing æ, ø, å with ae, oe, aa to examine orthographic effects.

Models. We evaluate four open-weight LLMs: LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Gemma-2-9B-it (Team et al., 2024), and Qwen2-7B-Instruct (Yang et al., 2024). None are explicitly fine-tuned on Danish and the extent of Danish in their pre-training corpora is unclear. In preliminary tests, we found that models exhibit some Danish comprehension but cannot consistently produce fluent Danish, suggesting limited exposure during pre-training. All models are fine-tuned using LoRA adapters (Hu et al., 2021) with rank 16 and alpha 32.

Evaluation Design. We examine LLM effectiveness across three dimensions: (1) prompting strategies, (2) language-specific factors, and (3) data conditions. For prompting, we study (RQ1) effectiveness of different strategies and (RQ2) cross-prompt generalization. For language-specific factors and data conditions we investigate (RQ3) Danish orthographic conventions and (RQ4) preprocessing strategies. All experiments use fine-tuning to evaluate performance under both balanced and imbalanced distributions, including few-shot settings.

We evaluate eleven prompting patterns in five categories: (1) *format-based*: vanilla_qa, choice_qa, cloze (Zhang et al., 2024b); (2) *reasoning-oriented*: chain-of-thought (Wei et al., 2023) and target identification; (3) *theory-driven*: illocutionary (Austin, 1962) and functional prompts; (4) *criteria-based*: explicit definitions and rules (Sigurbergsson and Derczynski, 2020); (5) *perspective-based*: victim viewpoints (Chiril et al., 2022) and expert moderator personas. Complete templates appear in Appendix A.

3 Results

We evaluate models under three experimental configurations. Configuration A trains and tests on balanced data. Configuration B trains and tests on imbalanced data. Configuration C trains on balanced data and tests on imbalanced data to assess transfer robustness. The balanced dataset contains 850 samples with equal class representation. The imbalanced dataset preserves the natural distribution of the DKHate corpus with 3,289 samples and approximately 13% (425 samples) hate speech. Configuration C proved ineffective due to the model overfitting to the majority class, so our primary research focuses on Configurations A and B. Configuration C, transfer precision-recall distributions appear in Appendix F.

3.1 Results of Primary Classification

Table 1 reports zero-shot SFT results and Table 2 reports 8-shot SFT results across all eleven prompting patterns. Base model results without fine-tuning appear in Appendix B. From these results, we derive findings concerning (1) fine-tuning effectiveness under different data distributions and (2) sensitivity to prompting strategies.

First, open-weight LLMs demonstrate meaningful potential for Danish hate speech detection with minimal task-specific fine-tuning. Additional visualizations appear in Appendix F and D revealing base and fine-tuned models consistent improvements under both balanced and imbalanced settings suggesting these models were exposed to Danish during pre-training.

Under balanced evaluation, Gemma achieves the strongest results for both zero-shot and few-shot using format-based patterns. LLaMA and Qwen also

Pattern	Balanced				Imbalanced			
	Ll	Mi	Ge	Qw	Ll	Mi	Ge	Qw
vanilla	.55	.44	**.89**	.77	.27	.00	**.52**	.10
choice	.48	.44	**.77**	.67	.18	.56	**.62**	.19
cloze	.76	.66	**.87**	.66	.23	.14	**.47**	.20
cot	**.64**	.38	.44	.60	.19	.21	**.37**	.26
target	.66	.44	.68	**.77**	.29	.18	**.54**	.04
illocution.	.75	.44	.56	**.79**	.16	.34	**.36**	.26
functional	.50	.16	.65	**.74**	.43	.29	.51	**.56**
definition	.32	.66	.56	**.77**	.30	.32	**.61**	.22
victim	.73	.68	**.84**	.70	.27	.00	**.57**	.17
expert	**.76**	.57	.58	.32	**.35**	.28	.40	.24
rules	.39	.64	.64	**.82**	.26	.11	**.38**	.25

Table 1: F1 scores with zero-shot prompting (SFT). Ll=LLaMA, Mi=Mistral, Ge=Gemma, Qw=Qwen. Bold indicates best per row.

Pattern	Balanced				Imbalanced			
	Ll	Mi	Ge	Qw	Ll	Mi	Ge	Qw
vanilla	.13	.24	**.80**	.64	.36	.00	**.49**	.36
choice	.55	.00	**.86**	.68	.18	**.33**	.17	.11
cloze	**.77**	.62	.71	.66	.23	.04	.00	**.29**
cot	.70	.49	.71	**.77**	.34	.00	**.41**	.18
target	.61	.54	.55	**.73**	.08	.25	**.62**	.36
illocution.	.70	.00	**.83**	.73	.32	.22	**.54**	.49
functional	.05	.63	.60	**.79**	.46	.29	**.62**	.27
definition	.29	.24	.44	**.78**	.51	.20	**.60**	.19
victim	.54	.61	.55	**.77**	.00	.04	**.51**	.38
expert	**.84**	.63	.59	.78	**.39**	.00	.31	.31
rules	.43	.05	.46	**.71**	.39	.03	**.46**	.00

Table 2: F1 scores with 8-shot prompting (SFT). Ll=LLaMA, Mi=Mistral, Ge=Gemma, Qw=Qwen. Bold indicates best per row.

perform well with perspective- and theory-based prompts respectively. These results are encouraging given the limited fine-tuning data.

However, under imbalanced conditions (13% hate speech), performance degradation is observed. The strongest configuration is Gemma with a theory-based prompting approach that analyzes social function scoring F1 of 0.62 for zero-shot and few-shot. This shift suggests theory-informed prompts provide more robust cues under class skew.

Second, model performance is highly sensitive to prompting strategy (**RQ1**). This reflects both the lightweight fine-tuning applied in contrast with English-centric LLMs undergoing large-scale SFT and the fundamental role of prompt design with even same-pattern fine-tuning and evaluation showing substantial variation across patterns. Gemma performs best with format-based prompts, LLaMA with cloze and perspective patterns, while Qwen shows the least variation.

Mistral exhibits the most severe instability with failures on choice_qa, illocutionary, and rules (format-, theory-, and criteria-based respectively). These stem from potential weaker instruction-following as the model generates outputs deviating from expected formats. This persists across runs suggesting fundamental prompt sensitivity that lightweight fine-tuning cannot overcome on Danish language.

3.2 RQ2: Cross-Prompt Generalization

We evaluate whether representations learned during fine-tuning transfer across prompting patterns by training on one pattern and testing on all others. Complete transfer matrices appear in Appendix E. Cross-pattern generalization varies substantially across models. On imbalanced data, `victim_perspective` achieves highest mean cross-pattern F1 for Llama and Mistral, while `vanilla_qa` performs best for Gemma. On balanced data, `illocutionary` shows strong transfer for Llama and Qwen, while `expert_moderator` and `cot` excel for Gemma. Two consistent patterns emerge. First, format-constrained patterns `choice_qa` and `cloze` exhibit near-zero cross-pattern transfer across all models—rigid output structures do not generalize when evaluation prompts expect different formats. Second, semantically grounded patterns framing classification in terms of intent, impact, or perspective transfer more robustly than format-driven alternatives, though the best-performing pattern varies by model. Prompt selection for SFT should thus consider not only same-pattern performance but also deployment context.

3.3 RQ3: Danish Orthographic Effects

Table 3 presents orthographic effects. On imbalanced data, ASCII transliteration consistently degrades performance, with Gemma showing the largest drop (0.486 to 0.258). This contradicts the hypothesis that Danish characters might confuse tokenizers trained primarily on English—if so, ASCII transliteration should improve performance. Instead, consistent degradation demonstrates that multilingual pre-training has established meaningful representations for æ, ø, and å; replacing them destroys learned signal.

Qwen presents an exception, showing slight improvement with ASCII on imbalanced data, possibly reflecting tokenizer design differences. This finding aligns with Müller-Eberstein et al. (2025),

	LLaMA	Mistral	Gemma	Qwen
Balanced				
Original (æøå)	.594	.501	.679	**.691**
ASCII (aeoeaa)	**.597**	**.562**	.550	.579
Imbalanced				
Original (æøå)	**.267**	**.283**	**.486**	**.227**
ASCII (aeoeaa)	.187	.149	.258	.250

Table 3: F1 comparing original Danish orthography versus ASCII transliteration (SFT, averaged across patterns).

	LLaMA	Mistral	Gemma	Qwen
Balanced				
Standard	**.594**	.501	**.679**	**.691**
Denoising	.547	**.555**	.600	.582
Imbalanced				
Standard	**.267**	**.283**	**.486**	.227
Denoising	.220	.155	.263	**.261**

Table 4: F1 comparing standard versus denoising prompts (SFT, averaged across patterns).

who show native Danish data outperforms translated alternatives, suggesting cross-lingual normalization loses important information.

3.4 RQ4: Preprocessing Strategies

Table 4 compares standard prompts against denoising prompts that explicitly instruct models to ignore social media noise. Denoising provides no benefit and often degrades performance. On imbalanced data, Gemma drops from 0.486 to 0.263 with denoising prompts.

These results suggest that instruction-tuned LLMs already handle social media noise effectively without explicit guidance. The models were pretrained on diverse web text including informal language and textual noise. Adding denoising instructions may harm performance by shifting attention toward identifying and filtering noise rather than focusing on classification. The cognitive overhead of explicit noise handling appears to outweigh potential benefits.

4 Discussion

The disconnect between balanced and imbalanced evaluation is striking. Configurations achieving high F1 on balanced data collapse by half on imbalanced data. This aligns with work showing hate speech comprises only a small fraction of real content (Tonneau et al., 2025). Balanced evaluation alone provides fundamentally misleading deployment guidance. Gemma consistently outperforms other models despite identical fine-tuning, suggesting architectural choices matter more than fine-tuning strategy. The ASCII degradation demonstrates that multilingual pre-training has established meaningful representations for Danish characters; preprocessing pipelines should preserve original orthography. Zero-shot SFT often outperforms few-shot SFT. Zero-shot achieves the highest balanced F1 overall, and few-shot introduces additional instability with increased failures across models. This suggests that few-shot examples may interfere with fine-tuned representations rather than enhance them when working in Danish, and practitioners can achieve strong results with zero-shot fine-tuning alone.

Cross-prompt generalization reveals that format-constrained patterns consistently fail to transfer, while semantically-grounded patterns show more robust generalization. Though, the best-performing pattern varies by model. This suggests prompt selection should consider context, not just same-pattern performance.

5 Conclusion

We presented the first systematic evaluation of open-weight LLMs for Danish hate speech detection. Theory-driven prompts outperform format-based patterns under class imbalance, while format-constrained patterns fail to transfer across prompts. Zero-shot fine-tuning matches or exceeds few-shot performance while avoiding additional instability. Danish characters should be preserved as ASCII transliteration degrades performance. Most critically, strong balanced performance does not guarantee real-world readiness. We recommend complementing balanced evaluation with imbalanced evaluation to accurately estimate deployment performance.

Limitations

We evaluate only four models in the 7-9B parameter range without closed-source models or Danish-specific models from Danish Foundation Models (Enevoldsen et al., 2023). The performance gaps we observe may partly reflect the broader Danish NLP resource constraints documented by Zhang et al. (2024a). Additionally, the DKHate corpus provides only binary labels (offensive versus non-offensive), limiting analysis to coarse-grained clas-

sification. Future work should explore multi-class Danish hate speech detection distinguishing categories such as racism, sexism, and religious hatred, provided adequate samples exist for each class to ensure robust model learning across all categories.

Ethical Considerations

This work addresses hate speech detection, where false positives risk silencing legitimate speech while false negatives permit harmful content. Our finding that models achieve high recall but low precision on imbalanced data suggests deployment would disproportionately affect non-hateful content.

References

John Langshaw Austin. 1962. *How to Do Things with Words*. Oxford University Press, Oxford.

Patricia Chiril and 1 others. 2022. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, 14(1):322–352.

Kenneth Enevoldsen, Lasse Hansen, Dan S. Nielsen, Rasmus A. F. Egebæk, Søren V. Holm, Martin C. Nielsen, Martin Bernstorff, Rasmus Larsen, Peter B. Jørgensen, Malte Højmark-Bertelsen, Peter B. Vahlstrup, Per Møldrup-Dalum, and Kristoffer Nielbo. 2023. Danish foundation models. *Preprint*, arXiv:2311.07264.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Andreas Kirkedal, Barbara Plank, Leon Derczynski, and Natalie Schluter. 2019. The lacunae of Danish natural language processing. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 356–362, Turku, Finland. Linköping University Electronic Press.

Matteo Melis, Gabriella Lapesa, and Dennis Assenmacher. 2025. A modular taxonomy for hate speech definitions and its impact on zero-shot LLM classification performance.

Karsten Müller and Carlo Schwarz. 2020. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.

Max Müller-Eberstein, Mike Zhang, Elisa Bassignana, Peter Brunsgaard Trolle, and Rob van der Goot. 2025. Dakultur: Evaluating the cultural awareness of language models for danish with native speakers. *Preprint*, arXiv:2504.02403.

Huy Nghiem and Hal Daumé Iii. 2024. HateCOT: An explanation-enhanced dataset for generalizable offensive speech detection via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5938–5956, Miami, Florida, USA. Association for Computational Linguistics.

Bolette S. Pedersen, Nathalie Sørensen, Sanni Nimb, Dorte Haltrup Hansen, Sussi Olsen, and Ali Al-Laith. 2025. Evaluating LLM-generated explanations of metaphors – a culture-sensitive study of Danish. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 470–479, Tallinn, Estonia. University of Tartu Library.

Rigspolitiet. 2024. Hadforbrydelser i 2024: Rigspolitiets aarsrapport vedroerende hadforbrydelser. Technical report, The Danish National Police, Copenhagen, Denmark.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Manuel Tonneau, Diyi Liu, Niyati Malhotra, Scott A. Hale, Samuel P. Fraiberger, Victor Orozco-Olvera, and Paul Röttger. 2025. Hateday: Insights from a global hate speech dataset representative of a day on twitter. *Preprint*, arXiv:2411.15462.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *Preprint*, arXiv:2302.11382.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Mike Zhang, Max Müller-Eberstein, Elisa Bassignana, and Rob van der Goot. 2024a. Snakmodel: Lessons learned from training an open danish large language model. *Preprint*, arXiv:2412.12956.

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024b. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of LLMs in implicit hate speech detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12073–12086, Bangkok, Thailand. Association for Computational Linguistics.

A Prompt Templates

Pattern	Danish Template	English Translation	Response
vanilla_qa	Givet udsagnet, er dette udsagn hadefuld tale? Svaret skal indeholde Ja eller Nej. Udsagn: {text} Svar:	'Given the statement, is this statement hate speech? The answer must contain Yes or No. Statement: {text} Answer:'	ja / nej
choice_qa	Givet udsagnet, er dette udsagn hadefuld tale? A: Ja. B: Nej. Angiv venligst dit valg. Udsagn: {text} Svar:	'Given the statement, is this statement hate speech? A: Yes. B: No. Please indicate your choice. Statement: {text} Answer:'	a / b
cloze	Givet udsagnet, er dette udsagn hadefuldt eller neutralt? Udsagn: {text} Svar: Det er et [MASK] udsagn.	'Given the statement, is this statement hateful or neutral? Statement: {text} Answer: It is a [MASK] statement.'	hadefuldt / neutralt 'hateful / neutral'
cot	Er dette udsagn hadefuld tale? Svar Ja eller Nej, og giv derefter din forklaring. Udsagn: {text} Svar:	'Is this statement hate speech? Answer Yes or No, and then give your explanation. Statement: {text} Answer:'	ja / nej
target	Er dette udsagn hadefuld tale? Svar Ja eller Nej, og identificer derefter målgruppen. Udsagn: {text} Svar:	'Is this statement hate speech? Answer Yes or No, and then identify the target group. Statement: {text} Answer:'	ja / nej
illocution.	Analyser hvilken sproghandling dette udsagn udfører. Hadefuld tale fungerer typisk som udtryk for foragt. Er dette hadefuld tale? Udsagn: {text} Svar:	'Analyze which speech act this statement performs. Hate speech typically functions as an expression of contempt. Is this hate speech? Statement: {text} Answer:'	ja / nej
functional	Overvej tekstens sociale funktion. Udelukker den andre eller etablerer dominans? Er dette hadefuld tale? Tekst: {text} Svar:	'Consider the social function of the text. Does it exclude others or establish dominance? Is this hate speech? Text: {text} Answer:'	ja / nej
definition	Hadefuld tale angriber en person eller gruppe baseret på race, religion, køn, seksualitet eller handicap. Tekst: {text} Er teksten hadefuld? Svar:	'Hate speech attacks a person or group based on race, religion, gender, sexuality, or disability. Text: {text} Is the text hateful? Answer:'	ja / nej
victim_persp.	Forestil dig du er den person teksten handler om. Ville du føle dig angrebet? Er dette hadefuld tale? Tekst: {text} Svar:	'Imagine you are the person the text is about. Would you feel attacked? Is this hate speech? Text: {text} Answer:'	ja / nej
expert_mod.	Du er en erfaren moderator på sociale medier. Overtræder denne tekst retningslinjer mod hadefuld tale? Tekst: {text} Svar:	'You are an experienced social media moderator. Does this text violate guidelines against hate speech? Text: {text} Answer:'	ja / nej
rules	Hadefuld tale kræver angreb på en gruppe baseret på race, religion, køn eller nationalitet OG trusler eller nedværdigende sprog. Tekst: {text} Svar:	'Hate speech requires an attack on a group based on race, religion, gender, or nationality AND threats or degrading language. Text: {text} Answer:'	ja / nej

Table 5: Complete Danish prompt templates organized by category, with English translations.

B Base Model Few-Shot Results

	Balanced F1				Imbalanced F1			
Pattern	**LLaMA**	**Mistral**	**Gemma**	**Qwen**	**LLaMA**	**Mistral**	**Gemma**	**Qwen**
vanilla_qa	.590	.579	.556	.575	.252	.297	.284	.324
choice_qa	.577	.612	.560	.680	.261	.297	.332	.352
cloze	.527	.047	.275	.537	.227	.088	.222	.276
cot	.384	.523	.689	.583	.199	.282	.358	.320
target	.466	.590	.698	.582	.188	.311	.350	.331
illocutionary	.539	.651	.649	.644	.251	.302	.354	.333
functional	.447	.635	.464	.553	.177	.289	.252	.276
definition	.493	.653	.602	.426	.218	.282	.288	.223
victim_persp.	.517	.553	.588	.617	.222	.261	.331	.337
expert_mod.	.612	.582	.327	.627	.229	.274	.239	.346
rules	.444	.614	.649	.467	.176	.312	.315	.240
Mean	.509	.549	.551	.572	.218	.272	.302	.305

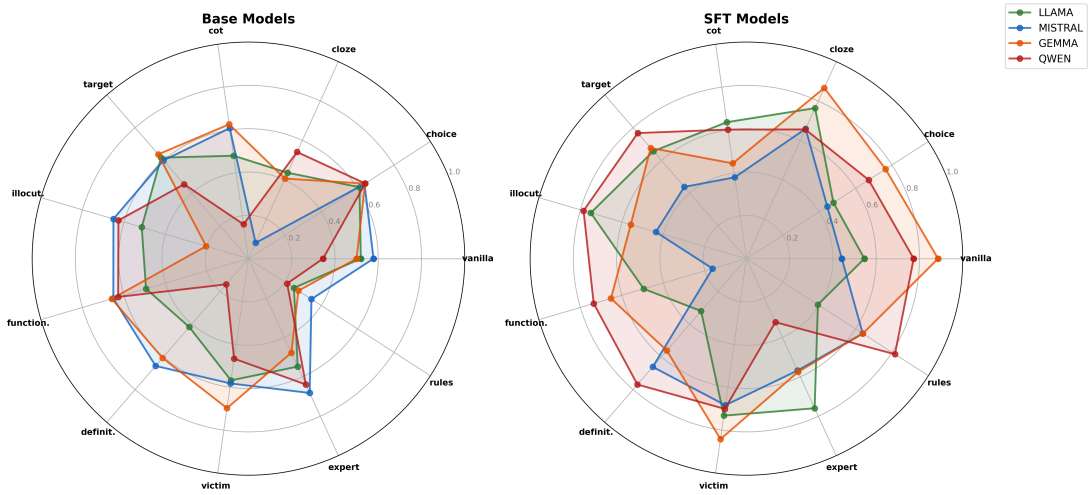
Table 6: Base model few-shot results without fine-tuning.

C Zero-Shot SFT Results

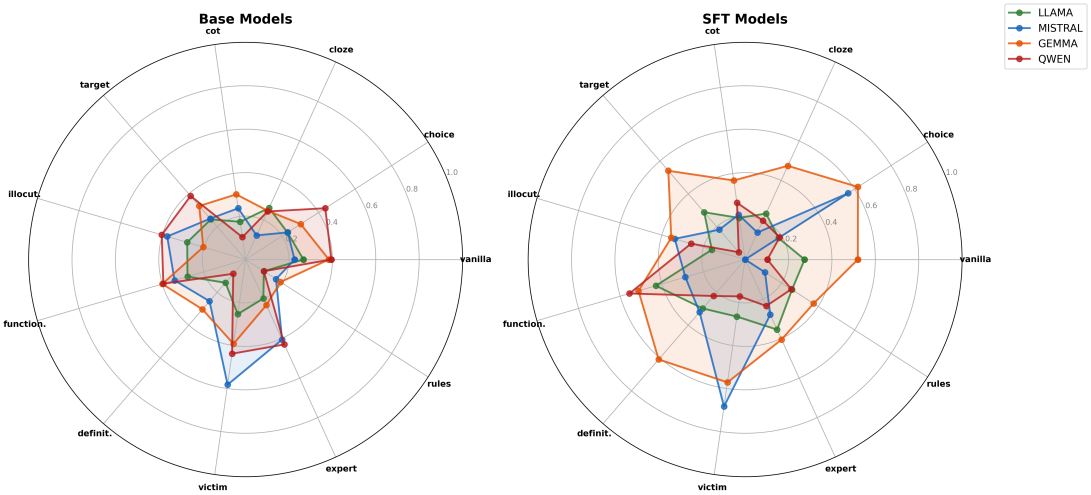
	Balanced F1				Imbalanced F1			
Pattern	**LLaMA**	**Mistral**	**Gemma**	**Qwen**	**LLaMA**	**Mistral**	**Gemma**	**Qwen**
vanilla_qa	.545	.441	**.886**	.773	.274	.000	**.519**	.103
choice_qa	.478	.444	**.765**	.672	.183	.564	**.618**	.189
cloze	.764	.659	**.867**	.656	.232	.136	**.474**	.197
cot	**.636**	.379	.444	.602	.194	.208	**.368**	.264
target	.657	.438	.676	**.767**	.288	.182	**.541**	.044
illocution.	.750	.436	.557	**.785**	.160	.338	**.355**	.260
functional	.495	.163	.653	**.736**	.429	.287	.512	**.556**
definition	.320	.661	.562	**.769**	.298	.321	**.609**	.222
victim_persp.	.732	.684	**.842**	.701	.265	—	**.571**	.171
expert_mod.	**.759**	.568	.575	.323	**.355**	.279	.405	.235
rules	.393	.639	.640	**.816**	.257	.109	**.375**	.253
Mean	.594	.501	.679	.691	.267	.242	.486	.227

Table 7: Zero-shot SFT results. — indicates corrupted data.

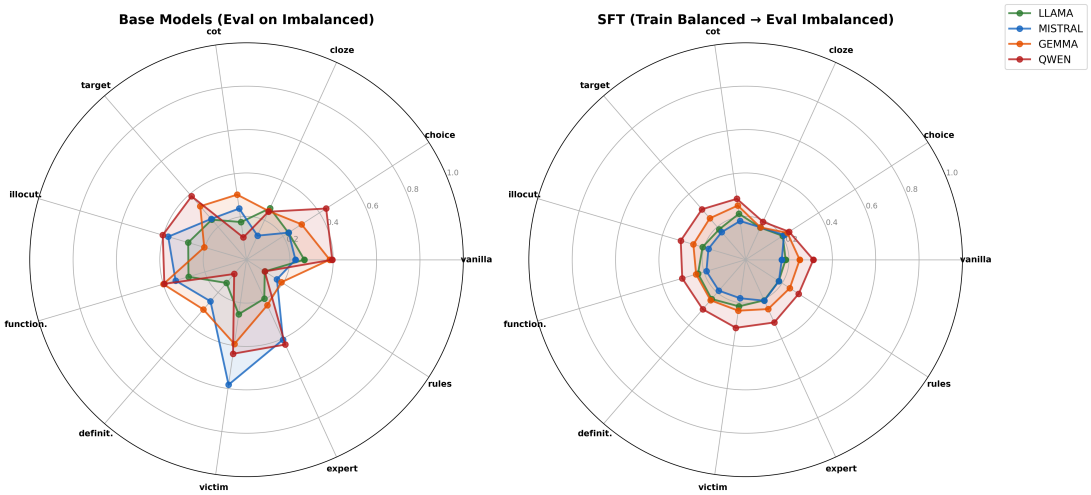
D Radar Charts by Evaluation Condition



(a) Balanced dataset performance across all eleven prompt patterns.



(b) Imbalanced dataset performance across all eleven prompt patterns.



(c) Balanced data and tested on imbalanced data across all eleven prompt patterns.

E Cross-Prompt Transfer Matrices

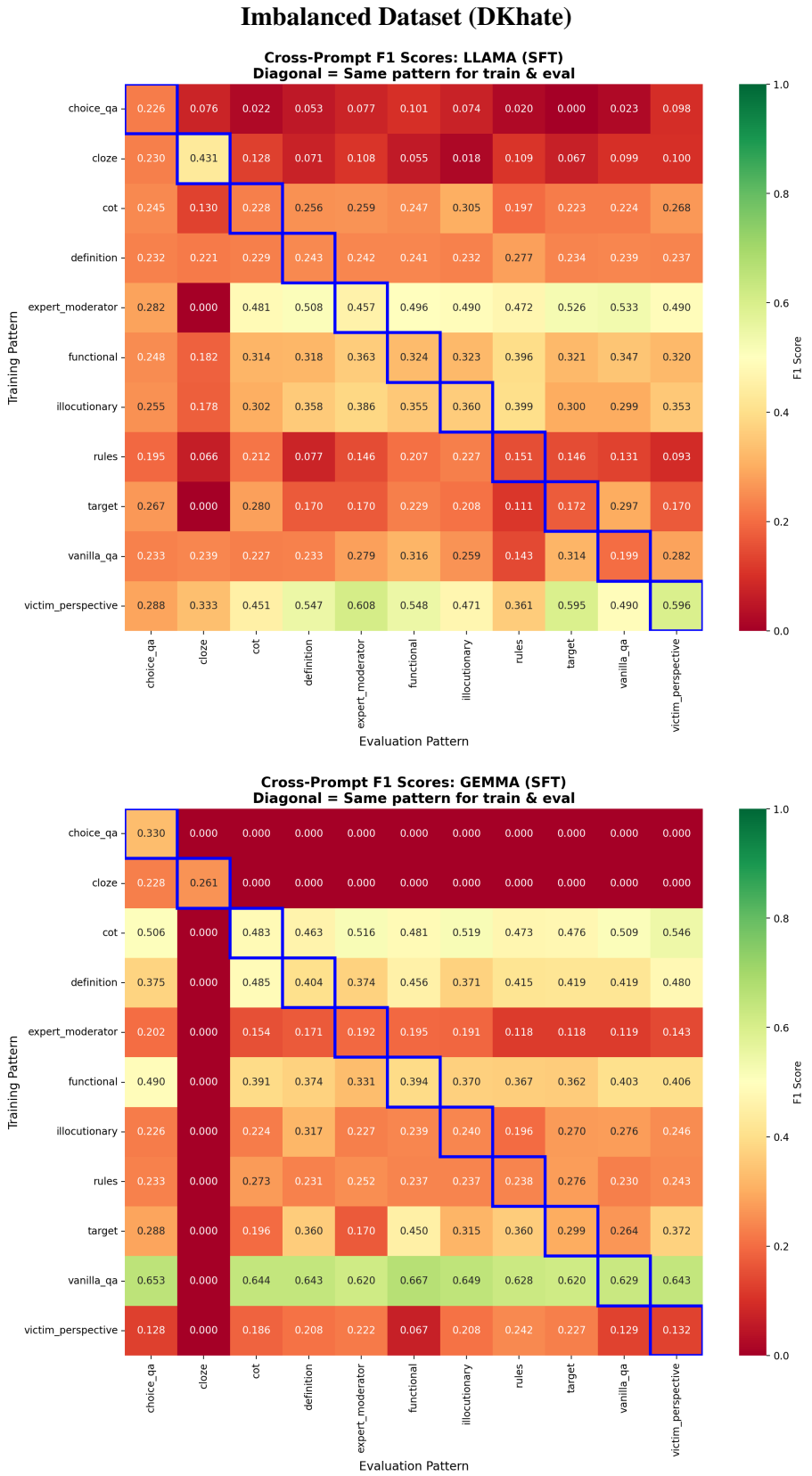


Figure 2: Cross-prompt transfer matrices (imbalanced dataset, part 1: LLaMA and Gemma).

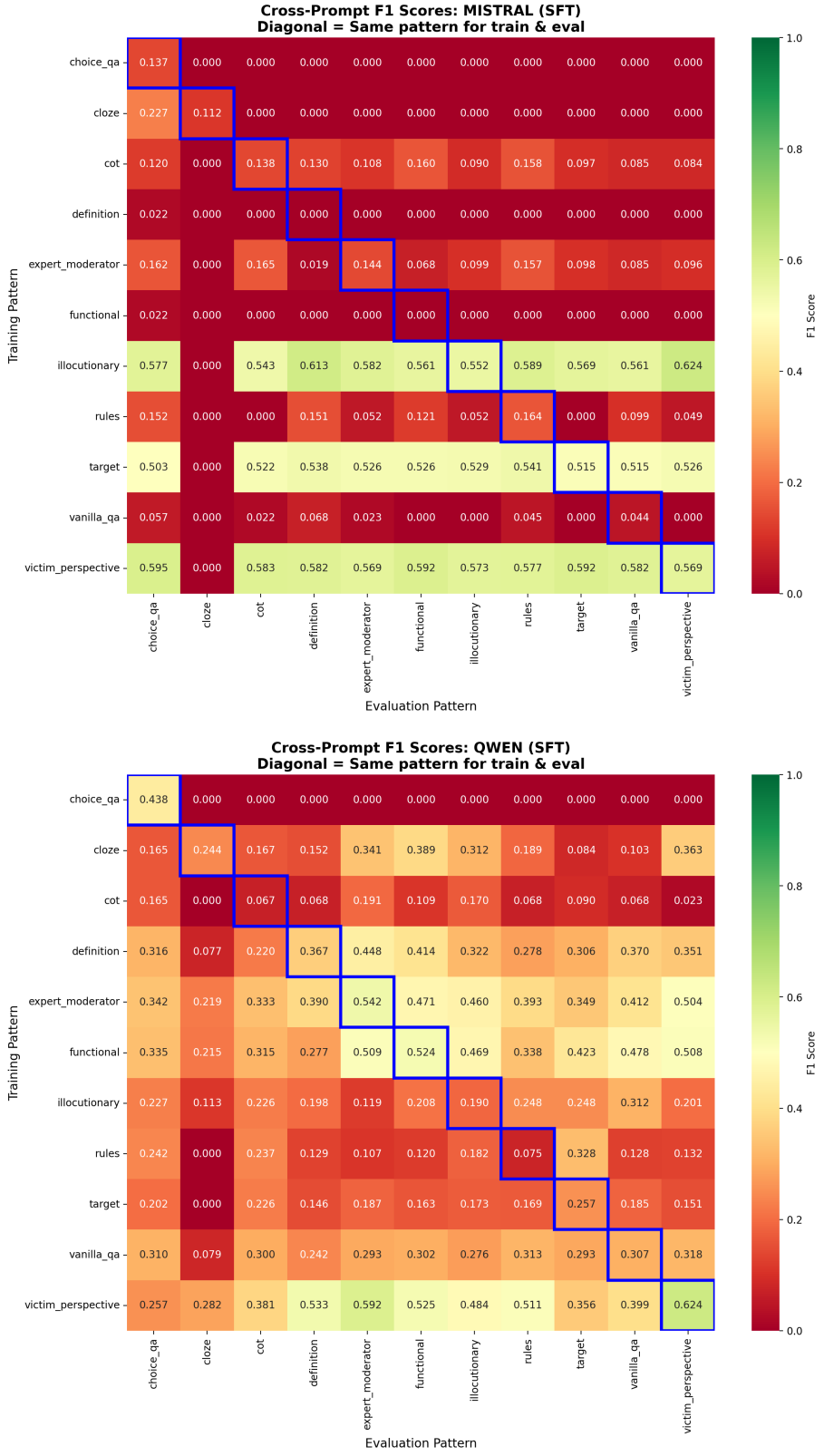


Figure 3: Cross-prompt transfer matrices (imbalanced dataset, part 2: Mistral and Qwen).

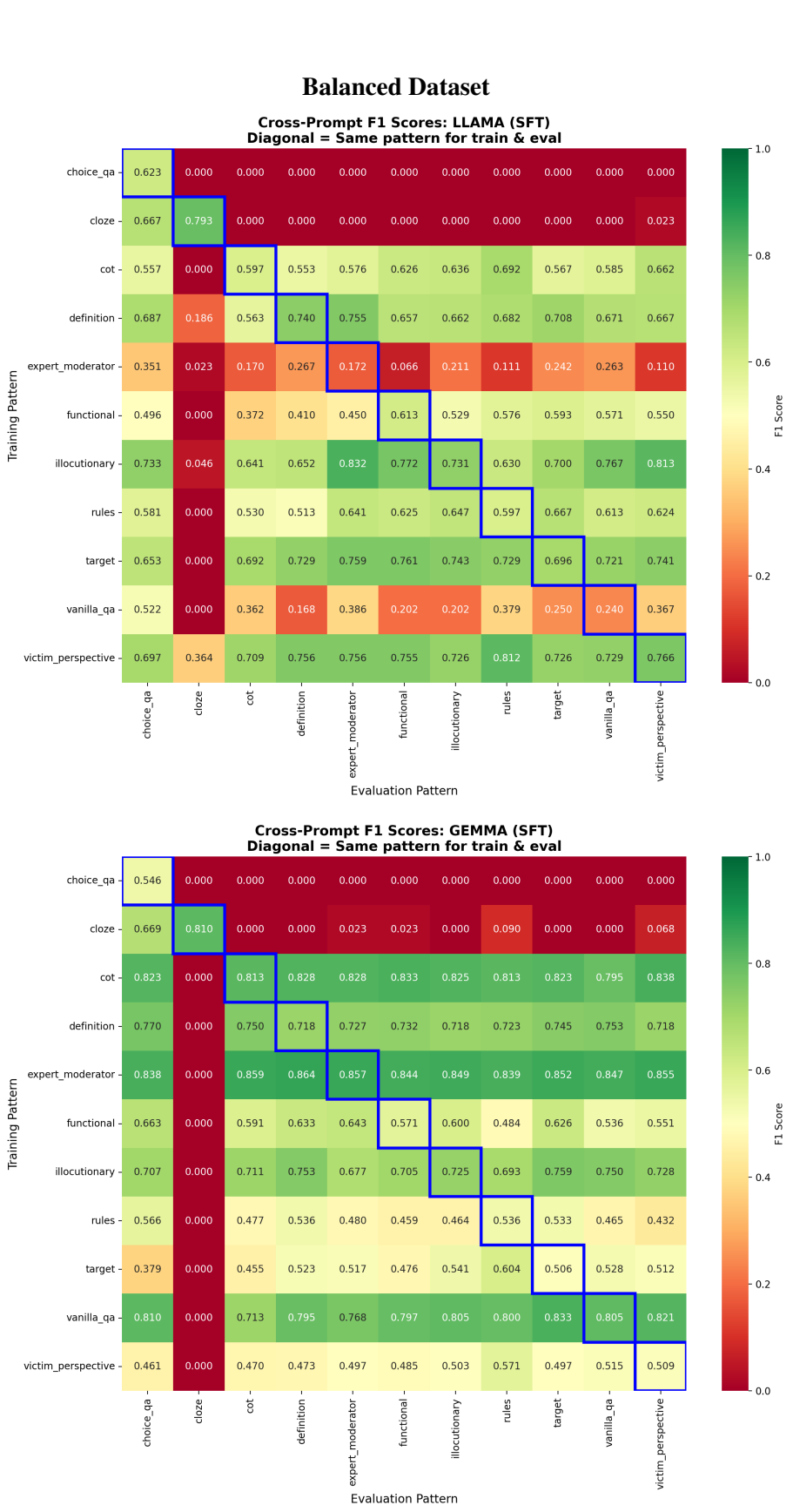


Figure 4: Cross-prompt transfer matrices (balanced dataset, part 1: LLaMA and Gemma).

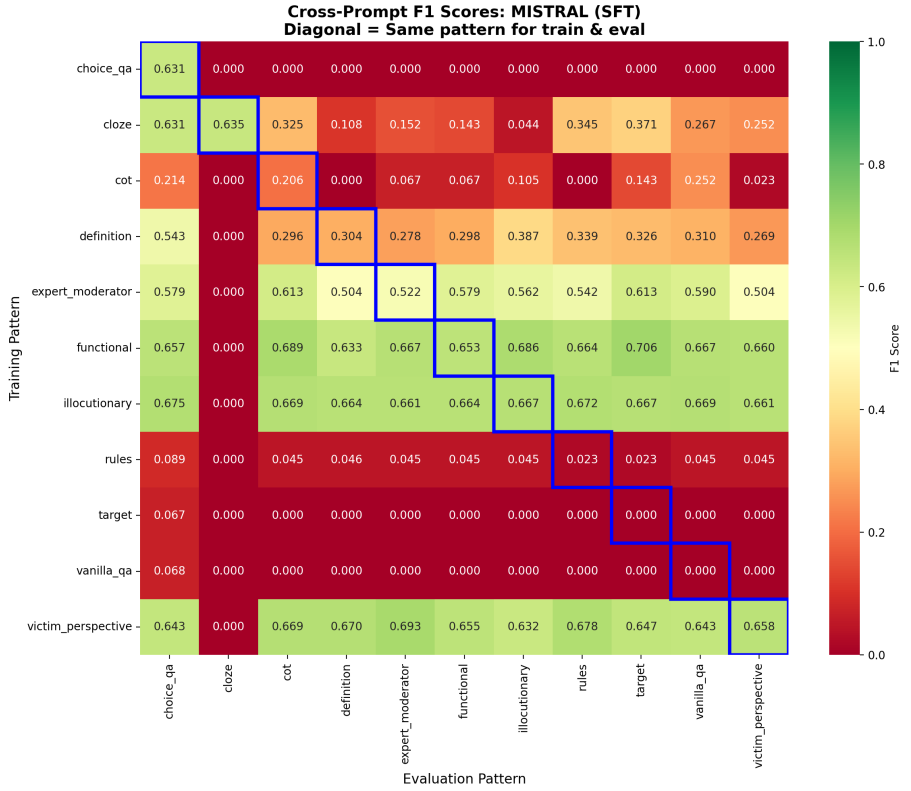


Figure 5: Cross-prompt transfer matrices (balanced dataset, part 2: Mistral and Qwen). Rows indicate training pattern; columns indicate evaluation pattern. Diagonal cells (blue outline) represent same-pattern performance; off-diagonal cells show cross-pattern generalization.

F Precision-Recall Distributions

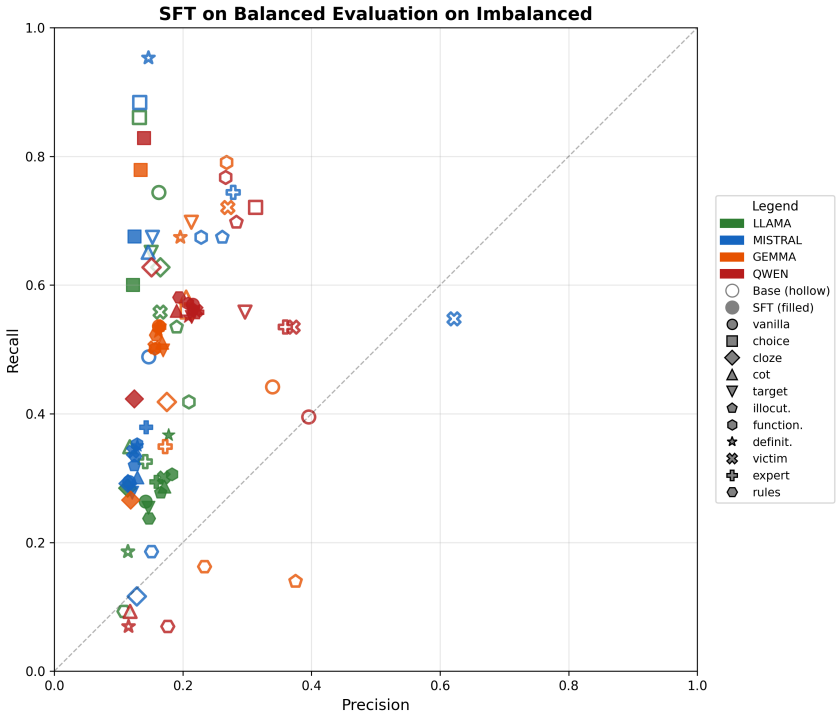


Figure 6: Precision-recall for distribution transfer (trained on balanced, tested on imbalanced).