hermeneutic_hools at SemEval-2026 Task 4: Multiperspectivity as a Resource for Narrative Similarity Prediction

Max Upravitelev[1,2], Veronika Solopova[1,2], Jing Yang[1,2,3], Charlott Jakob[1,2], Premtim Sahitaj[1,2], Ariana Sahitaj[1,2], and Vera Schmitt[1,2,3,4]
[1]Technische Universität Berlin
[2]German Research Center for Artificial Intelligence (DFKI)
[3]BIFOLD – Berlin Institute for the Foundations of Learning and Data
[4]Centre for European Research in Trusted AI (CERTAIN)
Correspondence: max.upravitelev@tu-berlin.de

Abstract

Predicting narrative similarity can be understood as an inherently interpretive task: different and equally valid readings of the same text can produce divergent interpretations and thus different similarity judgments, posing a fundamental challenge for semantic evaluation benchmarks that encode a single ground truth. Rather than treating this multiperspectivity as a challenge to overcome, we propose to incorporate it in the decision making process of predictive systems. To explore this strategy, we created an ensemble of 31 LLM personas. These range from practitioners following interpretive frameworks to more intuitive, lay-style characters. Our experiments were conducted on the SemEval-2026 Task 4 (Track A) dataset, where the system achieved an accuracy of 0.705 (leaderboard rank 12/46). Accuracy improves with ensemble size, consistent with Condorcet Jury Theorem-like dynamics under weakened independence. Practitioner personas perform worse individually but produce less correlated errors, yielding larger ensemble gains under majority voting. Our error analysis reveals a consistent negative association between gender-focused interpretive vocabulary and accuracy across all persona categories, suggesting either attention to dimensions not relevant for the benchmark or valid interpretations absent from the ground truth. This finding underscores the need for evaluation frameworks that account for interpretive plurality.

1 Introduction

Narrative understanding has traditionally been studied in interpretive sciences, where meaning is often modeled as perspective-dependent and shaped by subjective interpretation. As Kommers et al. (2025) emphasize, interpretations can conflict and even contradict each other while remaining valid, thus posing a fundamental challenge for computational approaches. This is especially relevant in the context of semantic evaluation and multiperspectivity: any benchmarks targeting interpretive tasks are subject to encoding the particular interpretive perspectives of their creators and annotators. Hence, predicting narrative similarity, recently formalized as SemEval-2026 Task 4 by Hatzel et al. (2026) and extending earlier works such as Hatzel and Biemann (2024); Akter and Santu (2024), is a task situated at the intersection of computation and interpretation.

In the following, we argue that multiperspectivity should be operationalized as a modeling component rather than treated as an evaluation artifact in narrative similarity prediction. This strategy is motivated by the Condorcet Jury Theorem (CJT), which states that collective decisions improve as independently competent voters are added (Shteingart et al., 2020). While its formal independence assumption is violated in LLM-based ensembles, the theorem motivates the hypothesis that aggregating diverse interpretive perspectives can improve prediction quality. Though this principle has yielded performance gains in NLP tasks like sentiment analysis (Bárcena-Ruiz and de Jesús Gil-Herrera, 2024), results across domains remain mixed (Lefort et al., 2024). Building on this observation, we construct an ensemble of 31 LLM personas divided into Practitioners [P] (personas with explicit analytical frameworks such as critical hermeneutics, feminist literary criticism or post-colonial theory) and Lay People [L] (more intuitive "everyday" characters). We systematically investigate how ensemble size, diversity, and persona category relate to prediction accuracy, and find that the relation between individual persona performance and collective decision quality reveals patterns about the role of interpretive diversity in computational narrative understanding.

Our error analysis further reveals a systematic pattern that bears directly on the relationship between multiperspectivity and semantic evaluation

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1817–1831

benchmarks: when personas across all categories employ vocabulary distinctive of gender-focused and feminist interpretation (terms such as *gender roles*, *female protagonist*, or *patriarchal*) their accuracy tends to drop. This finding allows for two complementary readings: Gender-related vocabulary may signal attention to narrative dimensions that, while interpretively valid, are simply not predictive of the similarity judgments encoded in the ground truth of the given benchmark. Alternatively, it may indicate that some perspectives produce valid but unrepresented interpretations, raising the question of how computational benchmarks should account for multiperspectivity in annotation and evaluation.

2 Preliminaries and Related Work

Narrative Similarity Computational approaches to narrative similarity span embedding-based methods for story reformulations (Hatzel and Biemann, 2024), structural decomposition via narrative theory (Chun, 2024), facet-based metrics grounded in 5W1H dimensions (Akter and Santu, 2024), and claim-level distillation for tracking ideas across media discourse (Waight et al., 2025). Narrative similarity also has applications in disinformation analysis, from taxonomy-based grouping of propaganda narratives (Nikolaidis et al., 2025) by similarity to comparing similar features across individual instances (Solopova et al., 2024).

Multi-agent Debating Systems While multi-agent debate (MAD) has become a prominent strategy for improving LLM reasoning and decision-making through iterative inter-agent interaction (Liang et al., 2024; Srivastava et al., 2025; Han et al., 2025), recent work has shown that majority voting alone often captures most of the associated performance gains across a range of knowledge and reasoning benchmarks (Choi et al., 2025; Kaesberg et al., 2025; Zhu et al., 2026). Building on this insight, we investigate multi-persona ensemble voting for narrative similarity assessment, while leaving the comparison with iterative debate protocols to future work. To the best of our knowledge, this is the first work to employ diverse LLM persona ensembles with majority voting grounded in distinct interpretive perspectives for narrative similarity prediction.

3 Methodology

Narrative Similarity Task Given a triplet of fictional texts in English consisting of an *Anchor* story as well as two candidate stories *A* and *B*, it should be decided which candidate is more similar to the anchor. The provided dataset is split into a development (dev) split with 200 samples and a test split with 400 samples, to be evaluated by measuring the accuracy of the predicted binary choice against gold data.

LLM Persona	Role	System Prompt
Literary Critic	Practitioner	You are a Literary Critic analyzing narrative similarity.
Computational Narratologist	Practitioner	You are a Computational Narratologist. You excel at structural analysis of text and the structured extraction of narrative features, such as actors, relations, events, topics, themes, sentiments etc.
...	...	...
Postcolonial Critic	Practitioner	You are a Postcolonial Critic examining the texts in terms of global power dynamics.
Feminist Literary Critic	Practitioner	You are a Feminist Literary Critic examining gender-related dynamics within texts.
Gender Perspective Analyst	Practitioner	You are a Gender Perspective Analyst examining gender-related dynamics.
...	...	...
High School Student	Lay	You are an 8th grader. You are not the best student, but you are doing alright.
Football Player	Lay	You are a professional football player and not sure why you got this task, but provide your perspective anyway.
...	...	...

Table 1: Selected LLM Personas, full list can be found in Appendix C

Crafting LLM Personas Since we want to investigate how the ensemble size and ensemble diversity relate to performance on the accuracy metric, we create 31 LLM personas and run different experiments on the dev set to investigate majority voting behavior in order to identify configurations that maximize accuracy on both the dev and the test set. Our ensemble consists of two categories of LLM personas, which we partly created manually and partly generated and refined: Practitioners [P], which are personas with concrete interpretive frameworks for handling text and text analysis, and Lay People [L], a group of more intuitive "everyday characters" inspired by findings from Kim et al. (2025), who showed that more complicated roles can lead to performance decreases when compared to simpler personas. Examples of our LLM personas can be found in Table 1. Every persona is essentially an LLM system prompt (like "You are

a Computational Narratologist. You excel at structural analysis of text and the structured extraction of narrative features, such as actors, relations, events, topics, themes, sentiments etc."), concatenated with base instructions (documented in Appendix G).

Expanding on the idea of multiperspectivity, we also deploy multiple models for each prediction and per persona. We chose three open-weight models from three different model families: Gemma 3 27-b it (Team et al., 2025), Qwen3-14B (Yang et al., 2025) and gpt-oss-20b (OpenAI et al., 2025).

Voting Configurations Each persona produces a single prediction per item on each model, yielding three levels of aggregation: (1) *Individual persona*: the prediction of one persona on one model, with no voting involved; (2) *Model-specific majority vote*: all 31 personas on a single model vote by simple majority; and (3) *Cross-model majority vote*: for each item, all 93 predictions produced by 31 personas $\times$ 3 models are pooled into a single majority vote, yielding one decision per item.

4 Results

Ensemble Size	Qwen3 14B	Gemma 3 27B-it	gpt-oss-20b	All
Majority Vote Accuracy				
E=1	69.1±0.4	71.9±0.4	57.0±0.7	66.0±0.3
E=3	71.7±0.5	73.4±0.4	58.4±0.9	69.8±0.4
E=5	72.7±0.6	73.8±0.5	59.0±0.9	71.5±0.4
E=10	73.8±0.7	73.9±0.4	59.8±1.1	72.9±0.6
E=20	74.2±0.7	74.2±0.5	60.2±1.4	74.5±0.6
E=30	74.4±0.8	74.4±0.7	60.5±1.6	75.0±0.7
E=31	74.3±0.9	74.5±0.8	60.8±1.9	75.2±0.6
Oracle $K \geq 1$ Accuracy				
E=1	68.6±0.4	71.9±0.4	56.4±0.7	65.6±0.3
E=3	87.5±0.4	84.9±0.4	84.6±0.6	90.4±0.3
E=5	92.3±0.4	88.7±0.4	92.5±0.5	95.8±0.2
E=10	96.3±0.3	92.9±0.5	97.8±0.3	98.8±0.1
E=20	98.4±0.3	95.9±0.8	99.5±0.2	99.8±0.1
E=30	99.1±0.4	97.1±1.0	99.8±0.2	99.9±0.1
E=31	99.1±0.4	97.2±1.0	99.8±0.2	99.9±0.1

Table 2: Ensemble accuracy (% ± standard deviation, based on random persona combinations with maximum sample limit of 5000.) by ensemble size, averaged over $n = 10$ runs.

Ensemble Size To investigate the effects of ensemble size, we run a series of experiments based on different size settings. The results are documented in Table 2. The upper block confirms that the accuracy of the majority vote increases with ensemble size E. The lower block reports an oracle analysis. Given access to the ground truth, we measure the proportion of items for which at least $K \geq 1$ members of the ensemble yield the correct prediction. Here, a similar behavior can be observed where the results plateau around $E = 30$.

Metric	Qwen3 14B	Gemma 3 27B-it	gpt-oss-20b
$K \geq 1$	$99.1\% \pm 0.4\%$	$97.2\% \pm 1.0\%$	$99.8\% \pm 0.2\%$
$K \geq 2$	$97.8\% \pm 0.5\%$	$94.5\% \pm 0.8\%$	$99.3\% \pm 0.5\%$
$K \geq 3$	$96.4\% \pm 0.5\%$	$92.5\% \pm 0.7\%$	$98.5\% \pm 0.7\%$
$K \geq 4$	$95.0\% \pm 0.5\%$	$90.6\% \pm 0.7\%$	$97.5\% \pm 0.7\%$
$K \geq 5$	$93.2\% \pm 0.6\%$	$88.6\% \pm 0.9\%$	$95.7\% \pm 0.9\%$
$K \geq 7$	$90.2\% \pm 0.9\%$	$84.7\% \pm 0.6\%$	$91.4\% \pm 1.3\%$
$K \geq 10$	$84.5\% \pm 0.6\%$	$80.4\% \pm 0.9\%$	$82.8\% \pm 1.7\%$
$K \geq 15$	$75.3\% \pm 0.5\%$	$75.4\% \pm 0.7\%$	$63.5\% \pm 1.7\%$
Majority	$74.0\% \pm 0.9\%$	$74.4\% \pm 0.7\%$	$60.1\% \pm 2.0\%$

Table 3: Cross-model comparison of K correct personas out of 31 in ensemble voting (mean and $\pm$ std across runs)

Oracle cross-model comparison The distribution of correct persona counts across samples is further illustrated in Table 3 for a cross-model comparison. As expected, the models follow different voting behaviors. For example, while gpt-oss-20b achieves the highest oracle accuracy at relaxed thresholds (up to $K \geq 7$), its performance drops sharply at stricter thresholds and falls well behind the other models at the majority vote level. This suggests that gpt-oss-20b exhibits high per-sample oracle coverage but low inter-persona agreement: individual personas frequently arrive at the correct answer, yet they rarely converge on it collectively. This pattern is unlikely to reflect model size alone: Qwen3-14B is smaller than gpt-oss-20b yet matches Gemma 3 27B-it on the majority vote (Table 4). One indicator is that individual per-persona accuracy on gpt-oss-20b averages $\approx 57\%$, compared with $\approx 69\%$ on Qwen3-14B and $\approx 72\%$ on Gemma 3 27B-it (Table 4); CJT-style majority-vote gains are bounded above by how far individual competence exceeds 0.5, so lower ensemble accuracy follows directly from lower individual accuracy. However, whether this lower individual accuracy reflects training-data composition or weaker instruction-following cannot be separated within this experimental setup.

Table 4 showcases the results across these configurations. Model-specific and cross-model majority votes consistently outperform individual persona predictions, with one exception: the "High School Student" persona matches the model-specific majority vote accuracy on Gemma 27B-it, though it is outperformed by the cross-model majority vote. The performance also differs across persona categories: [L] personas dominate the top 5 individual results while [P] personas occupy the bottom 5. In contrast, majority voting gives [P] an advantage

Rank	Persona	All	Qwen3	Gemma	gpt-oss
–	**MAJORITY VOTES ALL**	**75.8**	**74.3**	**74.5**	**60.8**
–	**MAJORITY [P] only**	**76.0**	**74.2**	**74.1**	**60.2**
–	**MAJORITY [P] (subsamp=13)**	**75.8**	**74.1**	**74.1**	**60.4**
–	**MAJORITY [L] only**	**75.3**	**73.5**	**74.5**	**59.1**
1	[L] Tour Guide	67.70	71.70	73.60	57.60
2	[P] Activist	67.20	70.30	74.20	57.10
3	[L] Small Business Owner	67.10	70.20	74.20	56.90
4	[L] Football Player	67.00	69.30	74.00	57.80
5	[L] Taxi Driver	67.00	70.90	74.40	55.70
6	[P] Storyteller	67.00	69.60	72.80	58.40
7	[P] Concise Reader	66.90	68.70	73.50	58.30
8	[P] Data Scientist	66.70	69.20	73.70	57.20
9	[P] Computational Narratologist	66.70	68.60	72.50	59.00
10	[L] Barkeeper	66.70	69.30	74.20	56.50
11	[L] High School Student	66.60	68.20	74.50	57.10
12	[L] Fitness Coach	66.60	68.90	72.00	58.80
13	[P] Literary Critic	66.60	69.10	72.00	58.50
14	[P] Computer Scientist	66.50	69.40	71.80	58.30
15	[L] Construction Worker	66.40	70.10	72.40	56.60
16	[L] Nurse	66.40	69.70	73.30	56.00
17	[P] Hermeneutics Specialist	66.20	69.90	72.40	56.10
18	[P] Journalist	66.10	69.50	71.20	57.50
19	[P] Critical Theorist	66.10	70.90	69.00	58.30
20	[L] Retiree	66.10	68.20	72.90	57.00
21	[L] Pirate	66.00	70.30	72.40	55.40
22	[L] Electrician	65.90	69.40	70.60	57.50
23	[P] Ethicist	65.80	70.40	70.90	56.10
24	[L] Line Cook	65.70	68.80	72.50	55.80
25	[P] Psychologist	65.70	67.00	71.80	58.00
26	[P] Yellow Press Journalist	65.60	69.60	72.30	54.90
27	[P] Fact Checker	65.20	68.60	69.20	57.70
28	[P] Psychoanalyst	64.90	69.10	68.20	57.30
29	[P] Gender Perspective Analyst	64.40	67.30	69.70	56.10
30	[P] Postcolonial Critic	64.30	70.20	67.00	55.70
31	[P] Feminist Literature Critic	63.60	67.80	66.20	56.70

Table 4: Per-Persona Accuracy (%) Comparison Across Models, sorted by the "All" column. [L] is short for Lay people, [P] for Practitioners.

over [L]. Since the category sizes are imbalanced (18 [P] vs. 13 [L]), we also perform a subsampling analysis where we randomly choose 13 [P] personas over 500 iterations, achieving a mean accuracy of 75.8% ($\sigma = 0.3\%$, 95% CI [75.2%, 76.4%]), comparable to the full [P] set (76.0%).

Further Results On the test set, the simple majority vote achieved 0.7025 accuracy. Our Track A leaderboard submission uses an optimized k=8 persona subset selected via exhaustive search with 5-fold cross-validation on dev (details in Appendix D); this subset reached $82.0\% \pm 4.8\%$ on dev but only 0.705 on test, which is nearly identical to the unsubmitted simple-majority baseline. The 11.5 percentage point dev-test gap (vs. 5.6 for the full ensemble) suggests overfitting to the small dev set (200 items), indicating that broad ensemble aggregation is more robust than targeted subset optimization for this task.

5 Error Analysis

To further investigate the performance split between [P] and [L] found in Table 4 and to specifically analyze why the "Gender Perspective Analyst" and "Feminist Literary Critic" (grouped together as $[P_G]$) personas particularly underperformed, we run a series of error and statistical analyses. We choose these personas because of their low performance, their similarity to each other (compared to the other personas) and because feminist literary criticism is among the most established approaches in literary studies (Plain and Sellers, 2007; Cooke, 2020). Given the rich theoretical and methodological foundation, the underperformance of these two personas is unexpected and warrants closer investigation if gender-analytical framing might interfere with narrative similarity judgments.

Metric	[P]	[L]	Conclusion
Individual accuracy	71.0%	71.7%	Higher accuracy for [L]
Error correlation (r)	0.388	0.461	Errors are 19% less correlated for [P]
Double-fault P(both wrong)	0.164	0.173	[L] fail together more often

Table 5: Error diversity analysis: comparison of [P] and [L] groups. Individual metrics are averaged across personas within each category; diversity metrics are averaged across all pairwise persona combinations within each category. Following (Kuncheva and Whitaker, 2003).

Error Diversity Analysis A consistency analysis across persona ensembles shows that $[P_G]$ are among the least consistent voters, while ensemble disagreement proxies item difficulty, with high-agreement items corresponding to clearer narrative similarity cases (see Appendix A). Table 4 reveals that [P] personas generally perform worse individually yet achieve a higher majority vote accuracy. To investigate this behavior, we perform a pairwise error correlation analysis considering individual LLM persona outputs and group performance. Following (Kuncheva and Whitaker, 2003), we compute pairwise error correlation and double-fault rate across personas within each category. The results in Table 5 point at one key difference between the groups: Members of [L] vote more unanimously, while votes from members of [P] with more concretely formulated perspectives vote more distinctly (further details in Appendix E).

Gender/Feminist vocabulary correlation To investigate whether gender-focused interpretation affects accuracy, we collected vocabulary distinctive to $[P_G]$ from LLM outputs (including explanations, themes and key points) using a lift-based approach:

Direction	Term	r_{pb}	p_{FDR}	Acc (present)	Acc (absent)	N
Negative correlations (gender vocabulary associated with lower accuracy)						
↓	female	−0.046	<0.001	56.5%	66.6%	4,575
↓	gender roles	−0.039	<0.001	42.0%	66.2%	571
↓	objectification	−0.032	<0.001	13.6%	66.1%	81
↓	female protagonist	−0.030	<0.001	54.9%	66.3%	1,516
↓	sexual violence	−0.029	<0.001	34.9%	66.2%	186
↓	female character	−0.027	<0.001	47.4%	66.2%	430
↓	experiences husband	−0.024	<0.001	0.0%	66.1%	28
↓	gender	−0.022	<0.001	56.3%	66.2%	1,102
↓	traditional gender	−0.022	<0.001	24.6%	66.1%	61
↓	patriarchal	−0.021	<0.001	49.3%	66.1%	337
Positive correlations (gender vocabulary associated with higher accuracy)						
↑	agency	+0.019	<0.001	71.7%	65.9%	2,547
↑	fulfillment outside	+0.012	0.005	96.8%	66.1%	31
↑	individuals babies	+0.012	0.005	100.0%	66.1%	25
↑	physical danger	+0.011	0.008	89.6%	66.1%	48
↑	women within	+0.011	0.009	87.5%	66.1%	56

Table 6: Point-biserial correlations between presence of gender/feminist-distinctive vocabulary and voting accuracy for *Other Expert* personas (combined across all three models, all output fields). Terms were identified via lift-based analysis (lift ≥ 5.0) from the two gender/feminist persona outputs. Correlations are FDR-corrected (Benjamini–Hochberg, $\alpha = 0.05$). N = number of responses containing the term. Full results for all persona groups and per-model breakdowns are provided in Appendix F.

for each term, we computed the ratio of its relative frequency in [P_G] outputs to its relative frequency across all persona outputs, retaining terms with lift ≥ 5.0 as characteristic of gender-focused interpretation. We then measured point-biserial correlations between the presence of these terms and voting correctness across all persona groups (Table 6). The results reveal a predominantly negative association: when other members of [P] use gender-distinctive terms, their accuracy drops compared to responses where these terms are absent. While statistically significant after FDR correction, these correlations are small in magnitude ($|r_{pb}| \leq 0.046$), reflecting the large sample sizes involved; the pattern is therefore best understood as a consistent directional tendency rather than a strong predictive signal. It holds across all three models and extends to Lay personas as well as [P_G] themselves (Appendix F).

Illustrative example The negative correlation in Table 6 does not necessarily reflect misidentification of story content. For example, the item 107 is one of 24 items in the dev set where the ensemble majority is correct while [P_G] personas err with gender-distinctive vocabulary in their outputs. The anchor is a novel told from the perspective of a pacifist countess across four 19th-century wars; candidate A depicts a woman traveling to confront the prison system holding her husband; candidate B depicts a young soldier's encounters with a woman across the First World War. The gold label is B, also chosen by 74 of 93 cross-model votes. All four [P_G] outputs on Gemma 3 27B-it and Qwen3-14B pick A instead, since their extracted themes and key points correctly identify the female protagonist in both the anchor and A, and their rationales explicitly weight this parallel above the multi-period war-narrative structure that the anchor and B share.

6 Discussion

Our results highlight a specific relation between individual persona accuracy and collective decision quality. Table 5 can be read as an argument for interpretive diversity: while [P] personas individually underperform [L] personas, their more distinctly formulated perspectives produce less correlated errors (19% lower pairwise error correlation), which in turn yields a larger ensemble gain under majority voting (75.3% vs 76.0%). This aligns with a widely held intuition in ensemble learning that diversity of errors can compensate for lower individual accuracy under majority voting (Kuncheva and Whitaker, 2003), although the relationship between diversity measures and ensemble accuracy remains complex (Tekin et al., 2024). While all personas exceed 50% accuracy (satisfying CJT's competence condition), the independence assumption is violated. Table 5 confirms substantial pairwise error correlations (r=.388 for [P], r=.461 for [L]), which can limit majority vote gains in LLM-based ensembles (Lefort et al., 2024). Nev-

ertheless, cross-model majority voting yields clear improvements over individual performance (75.8% vs. 67.7% maximum), suggesting that the combination of distinct persona framings and separate model families introduces sufficient diversity for meaningful gains under weakened independence.

The near-identical test performance of simple majority voting and the optimized k=8 ensemble (0.7025 vs. 0.705) reinforces this point from a practical perspective. Despite achieving 82.0% on the dev set through exhaustive subset search, the optimized ensemble might have been subject to overfitting and offered no meaningful advantage on the test split data. This suggests that for interpretive tasks where item-level difficulty and annotation variability are high, aggregating a large and diverse set of perspectives is more robust than selecting a small optimized subset. The finding also aligns with the broader argument of the paper: in settings where multiperspectivity is a feature rather than a source of noise, the diversity of interpretation may matter more than precision of selection.

The negative correlation between gender-analytical vocabulary and accuracy admits two complementary readings. Gender-focused framing may attend to thematic aspects that, while interpretively plausible, are not predictive of the similarity judgments encoded in the benchmark. Alternatively, these terms may reflect valid interpretations unrepresented in the ground truth. Either way, if valid interpretive perspectives can be penalized by standard accuracy metrics, the question of how benchmarks should account for multiperspectivity becomes pressing as NLP tasks move deeper into interpretive territory (Kommers et al., 2025).

Limitations

Several limitations of this study should be acknowledged. First, our experiments investigate majority voting and ensemble optimization but do not evaluate multi-agent debate (MAD) systems, in which personas iteratively discuss and revise their judgments. Our results therefore do not establish that simple voting outperforms deliberative approaches; rather, they characterize the behavior of independent, non-interacting ensembles. Evaluating whether structured interaction between personas yields additional gains remains an important direction for future work. Second, while our statistical analyses reveal significant correlations between vocabulary usage and accuracy, correlation does not equal causation. The presence of gender-analytical terms may co-occur with particular narrative properties of the input texts rather than causally driving incorrect predictions. Our discussion of these findings is itself interpretive, and alternative explanations remain plausible. Third, the 93-vote cross-model ensemble is computationally expensive: on a single NVIDIA H100 GPU with our (unoptimized) setup, predicting a single sample takes on average $\approx$ 56 seconds, placing the full configuration well beyond typical real-life deployment times. Finally, no conclusions should be drawn about the demographic or professional backgrounds of the dataset's annotators from our persona-level results. The observation that lay personas like the Tour Guide outperform the Feminist Literary Critic does not imply that the annotator pool was composed of individuals resembling the former more than the latter; it reflects properties of the LLM role-playing behavior and the specific benchmark, not of human annotation practices.

Acknowledgments

The work on this paper is performed in the scope of the projects "VeraXtract" (16IS24066) and "newspolygraph" (reference: 03RU2U151C) funded by the German Federal Ministry for Research, Technology and Aeronautics (BMFTR).

References

Mousumi Akter and Shubhra Kanti Karmaker Santu. 2024. Fans: a facet-based narrative similarity metric. *Preprint*, arXiv:2309.04823.

Gerardo Bárcena-Ruiz and Richard de Jesús Gil-Herrera. 2024. Application of condorcet's jury theorem for enhancing sentiment analysis performance using bert transformers: A case study for spanish. *Advances in Artificial Intelligence*, page 5.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Hyeong Kyu Choi, Xiaojin Zhu, and Yixuan Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*. Spotlight.

Jon Chun. 2024. AIStorySimilarity: Quantifying story similarity using narrative for search, IP infringement, and guided creativity. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 161–177, Miami, FL, USA. Association for Computational Linguistics.

Jennifer Cooke, editor. 2020. *The New Feminist Literary Studies*. Twenty-First-Century Critical Revisions. Cambridge University Press, Cambridge.

Chen Han, Wenzhen Zheng, and Xijin Tang. 2025. Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15114–15129, Suzhou, China. Association for Computational Linguistics.

Hans Ole Hatzel, Ekaterina Artemova, Haimo Stiemer, Evelyn Gius, and Chris Biemann. 2026. SemEval-2026 Task 4: Narrative similarity and narrative representation learning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. Association for Computational Linguistics.

Hans Ole Hatzel and Chris Biemann. 2024. Story embeddings — narrative-focused representations of fictional stories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5931–5943, Miami, Florida, USA. Association for Computational Linguistics.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11640–11671, Vienna, Austria. Association for Computational Linguistics.

Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2025. Persona is a double-edged sword: Rethinking the impact of role-play prompts in zero-shot reasoning tasks. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 848–862, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Cody Kommers, Ruth Ahnert, Maria Antoniak, Emmanouil Benetos, Steve Benford, Mercedes Bunz, Baptiste Caramiaux, Shauna Concannon, Martin Disley, James Dobson, Yali Du, Edgar Duéñez-Guzmán, Kerry Francksen, Evelyn Gius, Jonathan Gray, Ryan Heuser, Sarah Immel, Richard So, Sang Leigh, and 19 others. 2025. Computational hermeneutics: Evaluating generative ai as a cultural technology. *SSRN Electronic Journal*.

Ludmila I. Kuncheva and Christopher J. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Baptiste Lefort, Eric Benhamou, Jean-Jacques Ohana, Beatrice Guez, David Saltiel, and Thomas Jacquot. 2024. Examining independence in ensemble sentiment analysis: A study on the limits of large language models using the condorcet jury theorem. *Preprint*, arXiv:2409.00094.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9004–9017. Association for Computational Linguistics.

Nikolaos Nikolaidis, Nicolas Stefanovitch, Purificação Silvano, Dimitar Iliyanov Dimitrov, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ion Androutsopoulos, Preslav Nakov, Giovanni Da San Martino, and Jakub Piskorski. 2025. Polynarrative: A multilingual, multilabel, multi-domain dataset for narrative extraction from news articles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31323–31345, Vienna, Austria. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Gill Plain and Susan Sellers, editors. 2007. *A History of Feminist Literary Criticism*. Cambridge University Press, Cambridge.

Hanan Shteingart, Eran Marom, Igor Itkin, Gil Shabat, Michael Kolomenkin, Moshe Salhov, and Liran Katzir. 2020. Majority voting and the condorcet's jury theorem. *Preprint*, arXiv:2002.03153.

Veronika Solopova, Viktoriia Herman, Christoph Benzmüller, and Tim Landgraf. 2024. Check news in one click: NLP-empowered pro-kremlin propaganda detection. In *Proceedings of the 18th Conference of*

the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 44–51, St. Julians, Malta. Association for Computational Linguistics.

Gaurav Srivastava, Zhenyu Bi, Meng Lu, and Xuan Wang. 2025. DEBATE, TRAIN, EVOLVE: Self-evolution of language model reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32764–32810, Suzhou, China. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Selim Furkan Tekin, Fatih Ilhan, Tiansheng Huang, Sihao Hu, and Ling Liu. 2024. LLM-TOPLA: Efficient LLM ensemble by maximising diversity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11951–11966, Miami, Florida, USA. Association for Computational Linguistics.

Hannah Waight, Solomon Messing, Anton Shirikov, Margaret E. Roberts, Jonathan Nagler, Jason Greenfield, Megan A. Brown, Kevin Aslett, and Joshua A. Tucker. 2025. Quantifying narrative similarity across languages. *Sociological Methods & Research*, 54(3):933–983.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pages 23650–23678.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2026. Demystifying multi-agent debate: The role of confidence and diversity. *arXiv preprint arXiv:2601.19921*.

Appendix

A Consistency Analysis

Persona	All	Qwen3	Gemma	gpt-oss
Yellow Press Journalist	83.3	83.9	91.8	74.2
Fitness Coach	83.1	83.7	93.0	72.7
Computational Narratologist	83.1	83.3	92.1	73.7
Psychologist	82.9	83.5	93.1	72.2
Electrician	82.8	83.4	91.8	73.2
Ethicist	82.8	83.2	93.1	72.1
Critical Theorist	82.7	84.0	92.3	71.8
Small Business Owner	82.7	83.6	92.5	72.0
High School Student	82.6	83.3	91.5	73.0
Barkeeper	82.5	83.3	91.0	73.3
Tour Guide	82.5	83.8	91.8	72.0
Activist	82.5	83.0	93.0	71.6
Postcolonial Critic	82.5	83.7	92.2	71.5
Nurse	82.4	82.3	92.5	72.2
Concise Reader	82.3	83.2	91.2	72.5
Construction Worker	82.3	83.1	91.2	72.7
Retiree	82.3	83.3	90.9	72.6
Line Cook	82.2	82.9	91.7	72.1
Psychoanalyst	82.2	84.0	91.9	70.7
Taxi Driver	82.1	81.8	91.2	73.2
Data Scientist	82.0	80.9	91.0	74.0
Hermeneutics Specialist	82.0	84.4	90.5	71.0
Journalist	82.0	82.0	90.7	73.3
Literary Critic	81.7	81.8	91.5	71.9
Computer Scientist	81.6	82.2	90.7	71.9
Football Player	81.6	81.4	91.8	71.6
Pirate	81.5	80.2	92.5	71.8
Storyteller	81.4	81.3	91.2	71.8
Feminist Literature Critic	81.0	81.1	90.7	71.4
Gender Perspective Analyst	80.9	81.1	90.2	71.3
Fact Checker	80.6	80.7	89.1	72.2
Average	**82.2**	**82.7**	**91.6**	**72.3**

Table 7: Voting consistency (%) per persona and model. Consistency is defined as the fraction of runs matching the modal (most common) vote for each item, averaged across all 200 items. Each persona was evaluated 10 times per item; a consistency of 100% means the persona always gave the same answer across all runs, while 50% indicates a maximally split vote. Personas are sorted by combined consistency (descending).

To measure prediction certainty as a potential error indicator and to see if it correlates with accuracy, we opt for sampling consistency over verbalized (self-reported) confidence scores: we run each persona 10 times at temperature t=1 and measure the proportion of runs that yield the same predicted label. This follows works showing that consistency across stochastic samples provides a more reliable uncertainty signal than verbalized confidence, which tends to be poorly calibrated and systematically overconfident in instruction-tuned LLMs (Xiong et al., 2024; Manakul et al., 2023).

To empirically validate this choice, we compared both signals as accuracy predictors at all three analysis levels reported below. Across all levels and models, sampling consistency was the substantially stronger predictor: at the item level, consistency correlated with accuracy at $r = +0.53$ (combined), while verbalized confidence showed near-zero correlation ($r = +0.10$). At the within-persona level, consistency yielded significant positive correlations for all 31 personas across every model (mean $\bar{r} = +0.42$ combined), whereas verbalized confidence reached significance for only 21 of 31 personas with a mean $\bar{r}$ of just $+0.07$. These results confirm that, consistent with prior findings, sampling consistency is the more reliable uncertainty proxy for this setting.

As shown in table 7, $[P_G]$ predictions are indeed among the least consistent. To check if more consistent personas tend to be more accurate, we analyzed correlations at three levels (all p-values were FDR-corrected using the Benjamini-Hochberg procedure):

- **Across-persona level:** The Pearson correlation between mean consistency and mean accuracy was non-significant for individual models (Qwen-14B: $r = +0.21$, $p_{\text{adj}} = .34$; Gemma-27B: $r = +0.16$, $p_{\text{adj}} = .43$; OSS-20B: $r = -0.00$, $p_{\text{adj}} = .99$), but reached significance in the combined analysis ($r = +0.53$, $p_{\text{adj}} = .004$).

- **Within-persona level:** Per item, all 31 personas showed significant positive correlations ($p_{\text{adj}} < .05$) across every model, with mean $\bar{r}$ values of $+0.43$ (Qwen-14B), $+0.30$ (Gemma-27B), $+0.27$ (OSS-20B), and $+0.55$ (combined).

- **Item level:** Averaging consistency and accuracy across all personas per item yielded Pearson r values of $+0.52$ (Qwen-14B), $+0.44$ (Gemma-27B), $+0.37$ (OSS-20B), and $+0.61$ combined (all $p < .001$; Spearman $\rho = +0.75$ combined).

Together, these results indicate that while the personas themselves have little effect on the consistency-accuracy correlation, item difficulty is a dominant factor: items on which personas agree tend to be answered correctly, while items with split votes are disproportionately likely to be wrong. Thus, ensemble disagreement can serve as a proxy for item difficulty and, by extension, expected accuracy.

B Ensemble Agreement as Inter-Annotator Agreement

To examine whether interpretive framing meaningfully influences similarity judgments, we treat each persona as an annotator: for each item, we collapse the 10 stochastic runs into a single label via majority vote, then compute Krippendorff's α (nominal) over the resulting annotation matrix. Per individual model, agreement is substantial for Qwen 3-14B ($\alpha = 0.73$) and Gemma 27B ($\alpha = 0.72$) but only moderate for gpt-oss-20B ($\alpha = 0.49$). When all 93 persona × model combinations are treated as independent annotators, agreement drops to $\alpha = 0.42$ (moderate). Notably, [P] personas consistently show lower agreement than [L] personas (combined $\alpha = 0.40$ vs. 0.45; per-model differences of 6–13 percentage points), mirroring the lower pairwise error correlations reported in Table 5.

These values can be situated relative to the benchmark's own inter-annotator agreement: Hatzel et al. (2026) report $\alpha = 0.33$ for pairwise human annotations, noting that the dataset was filtered to retain only difficult cases. That the persona ensemble achieves comparable or higher agreement than human annotators ($\alpha = 0.42$ combined, up to 0.73 per model) while still falling short of near-perfect consensus reinforces that narrative similarity judgments involve genuine interpretive variability, which is a property of the task itself, not merely an artifact of LLM generation noise.

C LLM Personas

The full list of LLM personas is documented in the following:

LLM Persona	Role	System Prompt
Literary Critic	Practitioner	You are a Literary Critic analyzing narrative similarity.
Computational Narratologist	Practitioner	You are a Computational Narratologist. You excel at structural analysis of text and the structured extraction of narrative features, such as actors, relations, events, topics, themes, sentiments etc.
Psychologist	Practitioner	You are a Psychologist analyzing emotional and motivational similarity.
Psychoanalyst	Practitioner	You are a Psychoanalyst analyzing the text from the perspective of engaging with the unconscious and how it relates to society, focusing on authorial intent, symbolism and conflicts.
Storyteller	Practitioner	You are a Storyteller analyzing narrative flow and engagement.
Computer Scientist	Practitioner	You are a Computer Scientist analyzing structural consistency. You don't care much for literature, but you are able to talk about text.
Data Scientist	Practitioner	You are a Data Scientist checking consistency in the data.
Ethicist	Practitioner	You are an Ethicist Critic analyzing normative implications.
Critical Theorist	Practitioner	You are a Critical Theorist analyzing the texts in regard to its function within late-capitalist society.
Postcolonial Critic	Practitioner	You are a Postcolonial Critic examining the texts in terms of global power dynamics.
Feminist Literary Critic	Practitioner	You are a Feminist Literary Critic examining gender-related dynamics within texts.
Gender Perspective Analyst	Practitioner	You are a Gender Perspective Analyst examining gender-related dynamics.
Fact Checker	Practitioner	You are a Fact-Checker verifying claims.
Concise Reader	Practitioner	You are a Concise Reader giving very short summaries.
Hermeneutics Specialist	Practitioner	You are a specialist in hermeneutics in the tradition of Paul Ricœur.
Activist	Practitioner	You fight for a better world and are an expert in political communication.
Journalist	Practitioner	You are a Journalist able to capture the narrative of any given text quickly.
Yellow Press Journalist	Practitioner	You are a Yellow Press Journalist. You are a specialist in writing texts which get everyone's attention.
High School Student	Lay	You are an 8th grader. You are not the best student, but you are doing alright.
Football Player	Lay	You are a professional football player and not sure why you got this task, but provide your perspective anyway.
Pirate	Lay	You are a pirate from the 16th century. Please provide your perspective.
Barkeeper	Lay	You are a barkeeper and spend a lot of time talking to many people from all walks of life.
Construction Worker	Lay	You are a veteran construction worker. You value things that are built on a solid foundation.
Nurse	Lay	You are a registered nurse. You are empathetic but very grounded in reality.
Taxi Driver	Lay	You are a taxi driver who has driven thousands of miles and heard a million stories.
Line Cook	Lay	You are a line cook in a busy diner. You have a blunt, no-nonsense attitude.
Fitness Coach	Lay	You are a personal trainer and fitness coach. You focus on action, momentum, and results.
Electrician	Lay	You are an electrician. You look for the 'current' in a story—how things are connected.
Small Business Owner	Lay	You run a local hardware store. You are practical and budget-conscious.
Tour Guide	Lay	You are a local city tour guide. You explain complex histories in engaging ways.
Retiree	Lay	You are a retired office manager with a straightforward, common-sense approach.

Ensemble Size	Accuracy (%)	Std (%)
1	77.50	±7.60
7	81.50	±5.40
8	82.00	±4.80
9	81.50	±7.30
15	81.00	±5.60

Table 9: Exhaustive Search results: Ensemble size vs. accuracy (mean ± std).

D Exhaustive Search Results for Voting Ensemble Optimization

To find optimal ensemble sizes consisting of different voting patterns we performed exhaustive search over all $\binom{n}{k}$ possible k-member subsets from the top-n=31 candidates. Each candidate ensemble was evaluated using 5-fold cross-validation with simple majority voting. We selected the ensemble size by maximizing accuracy. When the number of combinations exceeded 50,000, we approximated exhaustive search via random sampling. Table 9 shows an evaluation of different ensemble sizes in relation to accuracy.

The resulting selection consisted of "Computer Scientist", "Critical Theorist", "Electrician", "Fitness Coach", "Hermeneutics Specialist" from Qwen3 14B and "Data Scientist", "Tour Guide", "Yellow Press Journalist" from Gemma 3 27B-it while notably no personas from gpt-oss-20B were included by this selection strategy.

E Extended Error Diversity Analysis

Table 10 presents our extended error diversity analysis which focuses on collected metrics per model. Individual metrics are averaged across personas within each category; diversity metrics are averaged across all pairwise persona combinations within each category. All diversity metrics are consistently lower for [P] across settings, indicating more diverse error patterns despite lower individual accuracy.

F Gender/Feminist Vocabulary Correlation Analysis

F.1 Method

We identified vocabulary distinctive to $[P_G]$ relative to the remaining 16 personas in [P] and 13 in [L] using a lift-based approach. For each term t (unigrams and bigrams, extracted via NLTK (Bird et al., 2009) after lowercasing, stopword removal, and filtering to alphabetic tokens of length > 1), we computed $\text{lift}(t) = (f_{\text{gender}}(t) + \epsilon)/(f_{\text{other}}(t) + \epsilon)$ with $\epsilon = 10^{-8}$. A term was classified as *gender/feminist-distinctive* if $\text{lift}(t) \geq 5.0$ and it occurred at least 10 times in gender persona outputs. We analyzed four output fields per persona: *themes*, *key points*, *explanation*, and *all* (their concatenation).

This procedure identified 229 (Qwen3-14B), 935 (Gemma-27B), and 216 (GPT-OSS-20B) distinctive terms on the *all* field, with 1,223 in the combined analysis. Representative terms range from demographic markers (*female protagonist*, *male lead*) and analytical concepts (*gender roles*, *patriarchal*, *agency*) to relational terms (*caregiving*, *male bonding*) and critical vocabulary (*objectification*, *gendered*).

For each distinctive term t and persona group $[P_G]$, $[P \setminus G]$, [L], we computed the point-biserial correlation $r_{pb}(t, G) = \text{corr}(\mathbf{1}[t \in \text{response}], \mathbf{1}[\text{vote} = \text{GT}])$, excluding ties and requiring ≥ 5 occurrences per group. Multiple testing was controlled via Benjamini–Hochberg FDR correction (Benjamini and Hochberg, 1995) at $\alpha = 0.05$, applied independently within each persona group $\times$ output field combination.

F.2 Summary of Significant Correlations

Table 11 provides an overview of the number and direction of significant correlations across all analysis conditions.

Across all conditions, negative correlations substantially outnumber positive ones. The pattern is consistent across persona groups: gender/feminist personas themselves, other experts, and lay personas all show predominantly negative associations between gender-distinctive vocabulary and accuracy. Gemma-27B yields the most significant correlations, consistent with its larger distinctive term set; GPT-OSS-20B shows very few, suggesting less differentiated gender-related vocabulary.

F.3 Top Correlations by Persona Group

Table 12 presents the strongest correlations for each persona group in the combined (all models, *all* fields) analysis.

F.4 Per-Model Consistency

The direction of the negative associations is consistent across models, though effect sizes vary. For the top negative-correlation terms among Other Expert personas (*all* field), both Qwen3-14B and Gemma-27B show significant accuracy drops when terms

Table 10: Error diversity analysis comparing [P] and [L] persona categories.

	Combined		Qwen3-14B		Gemma-3-27B		GPT-OSS-20B	
	[P]	[L]	[P]	[L]	[P]	[L]	[P]	[L]
Panel A: Individual & Ensemble Performance								
Number of personas	18	13	18	13	18	13	18	13
Mean individual acc. (%)	71.0	71.7	69.2	69.6	71.0	73.1	57.3	56.8
Accuracy std (%)	1.0	0.6	—	—	—	—	—	—
Accuracy range (%)	3.6	1.9	—	—	—	—	—	—
Majority vote acc. (%)	76.0	75.3	74.2	73.5	74.1	74.5	60.2	59.1
Ensemble gain† (%)	+5.0	+3.6	+5.0	+3.9	+3.1	+1.4	+2.9	+2.3
Panel B: Pairwise Error Diversity (lower = more diverse)								
Vote agreement	.748	.782	—	—	—	—	—	—
Pearson r (correctness)	.388	.461	.366	.406	.516	.620	.211	.236
Cohen's κ	.387	.461	.366	.406	.515	.620	.211	.236
Double-fault rate	.164	.173	.173	.177	.190	.194	.234	.244
$P(\text{both} \geq 1)$	.395	.443	—	—	—	—	—	—
Panel C: Expert–Lay Differences (Δ = Expert − Lay)								
Δ Individual acc. (%)	−0.7		−0.4		−2.1		+0.5	
Δ Majority vote acc. (%)	+0.7		+0.7		−0.4		+1.1	
Δ Pearson r	−.074		−.040		−.104		−.025	
Δ Cohen's κ	−.074		−.040		−.105		−.025	
Δ Double-fault rate	−.010		−.005		−.004		−.010	

†Ensemble gain = majority vote accuracy − mean individual accuracy.
Combined results aggregate persona votes across all three models per evaluation case.
Entries marked — were computed only for the combined setting.

are present: *female* (52.5% vs. 69.9% for Qwen; 54.8% vs. 72.3% for Gemma), *gender roles* (44.0% vs. 69.7%; 38.7% vs. 71.6%), and *gender* (59.6% vs. 69.6%; 53.1% vs. 71.6%). GPT-OSS-20B did not reach significance for any of these terms after FDR correction, consistent with its overall low number of significant correlations.

G Prompts Collection

G.1 Structured Generation

For LLM persona predictions, we concatenate system prompts with base instructions and structured output enforced by the vllm (Kwon et al., 2023) inference engine containing the following keys:

- Themes: 2-3 themes interpreted from the text
- Key points: The most important narrative points of the text
- Evidence: Textual snippets supporting the analysis
- Confidence: Scores (1-10) for each analysis
- Similarity: Similarity scores (1-10) for texts A and B relative to the anchor
- Explanation: A rationale explaining the similarity judgment

Pydantic specification for generating structured outputs:

```
from typing import List, Literal
from pydantic import BaseModel, Field

class Analysis(BaseModel):
    themes: List[str] = Field(...,
        description="2-3
                     themes or
                     interpretations")
    key_points: List[str] = Field(...,
        description="Key
                     narrative points")
    confidence: int = Field(..., ge=1,
        le=10,
        description="Confidence 1-10")
    evidence: List[str] = Field(
        default_factory=list,
        description="Evidence snippets")

class PersonaResponse(BaseModel):
    analysis_anchor: Analysis
    analysis_a: Analysis
    analysis_b: Analysis
    score_a: int = Field(..., ge=1,
        le=10)
    score_b: int = Field(..., ge=1,
        le=10)
    explanation: str = Field(...,
        description="Brief explanation
                    of similarity scores")
    rationale: str = Field("",
        description="Why A or B is more
        similar to anchor")

class JudgeResponse(BaseModel):
    final_decision: Literal["A",
        "B"] = Field(description="Which
            story is more similar to
            anchor")
```

Model	Field	Gender/Fem. (self)			Other Experts			Lay Personas		
		Sig.	+	−	Sig.	+	−	Sig.	+	−
Qwen3-14B	themes	4	0	4	11	1	10	10	0	10
	key_points	5	0	5	12	0	12	11	1	10
	explanation	6	0	6	22	0	22	19	0	19
	all	17	0	17	37	1	36	25	1	24
Gemma-27B	themes	96	19	77	79	15	64	40	4	36
	key_points	30	1	29	43	10	33	16	0	16
	explanation	41	1	40	41	4	37	23	1	22
	all	149	15	134	137	25	112	82	4	78
GPT-OSS-20B	themes	3	0	3	0	0	0	0	0	0
	key_points	5	3	2	6	2	4	4	3	1
	explanation	0	0	0	0	0	0	0	0	0
	all	1	0	1	7	5	2	3	2	1
Combined	themes	99	30	69	75	21	54	48	12	36
	key_points	19	4	15	35	5	30	19	1	18
	explanation	36	2	34	56	12	44	23	0	23
	all	132	27	105	119	26	93	63	3	60

Table 11: Number of significant correlations (FDR-corrected, $p < 0.05$) between gender/feminist-distinctive term presence and voting accuracy. *Sig.* = total significant; + = positive (term presence associated with higher accuracy); − = negative (term presence associated with lower accuracy).

```
    final_confidence: int = Field(ge=1,
        le=10, description="Decision
        confidence 1-10")
    explanation: str = Field(description=
        "Brief explanation of decision")
```

G.2 System Prompts

Example system prompt with instructions for structured output generation:

```
You are a Literary Critic analyzing
narrative similarity.

Rules:
1. Score similarity 1-10 (higher = more
    similar to anchor)
2. Confidence 1-10 (higher = more confident)
3. Provide 1-3 evidence snippets per story
4. Brief chain_of_thought (CoT) for each
    analysis
5. Output ONLY the valid JSON object, no
    other text
```

G.3 User Prompt

```
user_prompt = f"""Compare these stories:

ANCHOR STORY:
{anchor}

STORY A:
{text_a}

STORY B:
{text_b}

Analyze which story (A or B) is more similar
    to the Anchor story."""
```

Group	Dir.		Term	r_{pb}	Acc (+)	Acc (−)	N
Gender/Fem. (self)	↑	1	male friendship	+.043	91.3%	63.9%	69
		2	male relationships	+.036	82.5%	63.9%	103
		3	male bonding	+.036	83.5%	63.9%	91
	↓	1	male character	−.058	45.7%	64.4%	269
		2	female	−.057	59.6%	65.7%	3,312
		3	affairs political	−.050	0.0%	64.1%	17
		4	experiences husband	−.049	0.0%	64.1%	16
		5	closeted identity	−.047	0.0%	64.1%	15
Other Experts	↑	1	agency	+.019	71.7%	65.9%	2,547
		2	fulfillment outside	+.012	96.8%	66.1%	31
		3	individuals babies	+.012	100.0%	66.1%	25
	↓	1	female	−.046	56.5%	66.6%	4,575
		2	gender roles	−.039	42.0%	66.2%	571
		3	objectification	−.032	13.6%	66.1%	81
		4	female protagonist	−.030	54.9%	66.3%	1,516
		5	sexual violence	−.029	34.9%	66.2%	186
Lay Personas	↑	1	agency	+.014	72.2%	66.5%	1,016
		2	protagonist distracted	+.012	100.0%	66.5%	24
		3	woman ultimately	+.010	90.6%	66.5%	32
	↓	1	female	−.041	57.4%	67.0%	3,320
		2	sexual violence	−.039	20.5%	66.6%	122
		3	female protagonist	−.034	53.7%	66.8%	1,182
		4	gender roles	−.032	37.2%	66.6%	207
		5	objectification	−.027	6.1%	66.6%	33

Table 12: Top significant correlations (FDR-corrected, $p < 0.05$) between gender/feminist vocabulary and voting accuracy, combined across all models, *all* fields. Top 3 positive (↑) and top 5 negative (↓) correlations shown per persona group. All $p_{\text{FDR}} < 0.005$. Totals: Gender/Feminist (self): 132 significant (27+, 105−); Other Experts: 119 (26+, 93−); Lay: 63 (3+, 60−).