A Hybrid AI System for Evaluating Media Representation of Violence and Inequality

Zhan Liu[0000−0003−3367−3204]⋆ and Nicole Glassey Balet[0000−0003−3268−8375]

Media Innovation Lab
University of Applied Sciences and Arts Western Switzerland
HES-SO Valais-Wallis, Sierre, Switzerland
zhan.liu@hevs.ch
nicole.glassey@hevs.ch

Abstract. Media coverage of gender-based violence plays a critical role in shaping public understanding and policy, yet often perpetuates stereotypes and biases. We present a hybrid AI approach to analyze how French-language media represent gender-based violence. Combining rule-based Natural Language Processing (NLP) with Large Language Models (LLMs), the system applies expert-defined criteria across analytical categories, with each criterion assigned to the most effective method based on empirical performance. This strategy achieves 87.1% overall accuracy, surpassing previous models. GPT-4 led general performance (77.9%), while NLP delivered exceptional results in structural and language-sensitive categories. Our findings demonstrate that combining complementary AI techniques enables near-human accuracy in evaluating media narratives and contributes to advancing web-based text mining for socially relevant media analysis.

Keywords: Hybrid AI · Large language models · Natural language processing · Web text mining · Social bias in media representation · Gender-based violence.

1 Introduction

Media representation of gender remains a critical factor in shaping public perceptions and reinforcing—or challenging—societal norms. Despite increased attention to gender equality, media coverage of gender-based violence (GBV) continues to exhibit stereotypes, bias, and misrepresentation, necessitating rigorous scrutiny and systematic evaluation methodologies. As [4] demonstrate, news media consistently produce content reinforcing traditional gender hierarchies despite growing awareness of inequality issues.

According to [19], GBV affects roughly one in three women globally, ranging from psychological abuse to femicide. Media portrayals significantly influence public understanding of violence's causes, severity, and solutions. [11] highlight

⋆ Corresponding author.

that sensational coverage of isolated incidents often distorts reality and hinders effective policy responses.

Swiss media outlets differ markedly in their GBV framing, showing variable adherence to ethical guidelines from organizations such as the Council of Europe- [7] and [9]. Systematic implementation and monitoring remain limited, creating a clear opportunity for scalable technological solutions, precisely the challenge addressed in this research.

Current methods for evaluating gender representation suffer from three main limitations: (1) lack of comprehensive, standardized evaluation frameworks, (2) limited understanding of how different AI techniques handle culturally-specific gender narratives, and (3) absence of hybrid approaches integrating complementary strengths of traditional NLP and advanced LLM technologies.

To address these gaps, this study introduces a structured taxonomy refining analytical criteria, significantly broadening coverage of gender representation dimensions. It also compares four prominent large language models—GPT-4, Llama 3, Gemma 3, and DeepSeek R1—to elucidate trade-offs in performance and accessibility. Finally, the proposed hybrid analytical strategy optimizes the integration of NLP and LLM methods based on empirical effectiveness, substantially improving accuracy and closely approaching expert human assessments.

Results show no single AI approach excels across all evaluation dimensions. Traditional NLP methods outperform LLMs in structural analysis, while LLMs excel at identifying implicit bias and contextual framing. Among tested models, GPT-4 achieved the highest overall accuracy, though open-source alternatives like Gemma demonstrated competitive performance, suggesting viable pathways for more accessible implementations.

Collectively, these contributions yield a robust, hybrid AI framework, significantly enhancing analytical capabilities in evaluating media narratives on gender-based violence.

2 Related work

2.1 Gender stereotypes and media representation

A substantial body of research demonstrates that news media coverage of gender-based violence (GBV) often reinforces harmful stereotypes. For example, political discourse in news frequently assigns gendered blame [3]. Many narratives perpetuate victim-blaming tropes [10], and news reports often emphasize individual pathology in GBV cases rather than examining broader societal factors [5]. These patterns perpetuate the misconception that incidents of GBV are isolated occurrences rather than manifestations of systemic inequality. Such biases also extend beyond textual content. AI-generated news imagery, for instance, overrepresents men and frequently portrays women in passive roles [6], indicating that emerging digital media platforms replicate the same gender stereotypes found in traditional journalism.

Structural factors within the media industry may further contribute to biased representation. Women occupy only about one-quarter of top editorial roles in

news organizations, despite comprising roughly 40% of the journalism workforce [16]. This imbalance likely influences which voices and perspectives dominate coverage. These representational problems are deeply intertwined with organizational cultures and professional norms [9], implying that remedies must involve changes at both individual and institutional levels. However, despite extensive documentation of media biases, there is still no comprehensive framework to systematically evaluate the nuanced and implicit ways GBV is framed in media coverage.

2.2 AI applications in media analysis

The rise of artificial intelligence (AI) and NLP has opened new opportunities for large-scale media analysis, but these tools are not inherently neutral and often reflect or amplify biases from their training data [15], highlighting the need for socio-technical caution.

News organizations are increasingly deploying AI for various newsroom tasks [12], but this adoption often occurs without sufficient scrutiny of algorithmic biases or ethical implications. Interviews with journalists indicate that many newsrooms implement AI without fully understanding its limitations [17]—an oversight that can inadvertently perpetuate the inequalities these tools are meant to mitigate. As a result, most AI-driven media analysis systems fail to account for the nuanced cultural context of gender narratives, especially in multilingual or sensitive settings.

2.3 Challenges in automated content evaluation

Despite progress, significant challenges persist in the development of automated systems to evaluate media content for gender bias. Bias can infiltrate every stage of the AI pipeline, often in ways that are not transparent [14]. LLMs may further amplify these biases, acting as "stochastic parrots" that reproduce biased patterns without true understanding [2]. For example, ChatGPT has been observed to default to male pronouns for "programmer" and female for "nurse" even when instructed otherwise [18]. Such outcomes illustrate the limitations of LLMs in producing unbiased language and underscore the need for continual bias evaluation and mitigation.

Frameworks like the Council of Europe's guidelines [7] provide benchmarks for fair and sensitive reporting, but translating these principles into concrete, automatable criteria remains difficult. Furthermore, few approaches combine rule-based NLP with LLMs, so subtle framing biases often go undetected. Hence, methods that operationalize ethical standards and integrate NLP and LLM approaches are crucial for advancing automated content evaluation.

2.4 Evolution from our previous work

The current study substantially extends previous research [13], which introduced a preliminary NLP-LLM hybrid approach for evaluating media coverage

of gender-based violence. While that work demonstrated the potential of hybrid methods (76% accuracy across 27 evaluation criteria), several critical limitations remain.

First, our earlier framework used a simplistic binary assignment of criteria (NLP vs. GPT), resulting in suboptimal alignments. We now empirically assign each criterion to its most effective method, greatly improving performance alignment. Second, the prior study relied only on GPT-4. We now evaluate multiple LLMs (Llama 3-12B, Gemma 3-12B, DeepSeek-R1-14B), highlighting performance, cost, and accessibility trade-offs between commercial and open-source models. Third, the previous work focused mainly on explicit content. The current framework expands to 32 criteria in 9 categories to capture implicit bias, linguistic framing, and systemic patterns. Finally, our hybrid method reaches 87.1% overall accuracy, an 11-point gain over the prior 76%, demonstrating the effectiveness of this optimized approach.

2.5 Research gap and contributions

Based on this literature review, we identify three primary research gaps:

1. Most existing studies focus narrowly on particular aspects of gender bias without offering an integrated framework capable of holistic evaluation.
2. Comparative evaluations of emerging open-source LLMs for media analysis (especially in non-English contexts like Swiss French journalism) remain scarce.
3. Prior work typically uses either traditional NLP or LLMs; few employ a hybrid approach leveraging each method's strengths.

This paper addresses these gaps by introducing a comprehensive, taxonomically structured evaluation framework, benchmarking commercial and open-source LLMs alongside traditional NLP tools, and implementing a hybrid methodology designed to maximize reliability and accuracy in gender representation analysis.

3 Dataset description

3.1 Data collection and sources

To capture the evolving landscape of GBV representation in Swiss French-language media, we expanded our data collection approach both in scope and in methodology. Our previous study relied on a narrower selection of media sources and keyword-based retrieval; the current study adopts a more comprehensive, multi-source strategy integrating both mainstream and emerging platforms.

Our corpus was provided by research partners who collected articles from Swiss media archives and monitoring tools from gender equality organizations. This multi-source approach enhanced both the breadth and timeliness of our

corpus, allowing us to include dynamic content such as comment sections and associated multimedia.

Our dataset spans from 2022 to early 2025, encompassing three full years of news coverage. We collected 5,517 French-language articles from 25 Swiss media outlets. In addition to our previous sources, we added six prominent digital-native platforms: *M Le Média*, *Heidi.news*, *La Côte*, *Blue News*, *aJour*, and *Frapp*. These platforms bring editorial diversity and introduce alternative framing practices that enrich our analysis.

This expansion significantly improves upon our previous dataset (1,752 articles, 19 outlets, 2022), offering greater size, media diversity, and temporal coverage for identifying longitudinal patterns in GBV representation.

3.2 Data preprocessing and refinement

We implemented a multi-phase preprocessing pipeline to ensure the dataset's quality and relevance, building on earlier improvements in semantic filtering and normalization.

- Semantic filtering: We augmented keyword matching with word embeddings and semantic networks to capture thematically related content, improving recall for GBV content expressed implicitly or in varied language.
- Relevance scoring and review: Each article received an automated relevance score, with low-scoring ones flagged for manual review to eliminate false positives.
- Metadata extraction: Using source-specific parsers, we standardized article metadata into 14 structured fields (e.g., date, source, section, tags) and noted whether articles contained images or videos for potential future analysis.
- Deduplication: We removed exact and near-duplicate articles (e.g., syndicated content with minor edits) using a similarity detection algorithm.

This preprocessing yielded 3,684 unique and relevant articles with complete metadata. Article lengths ranged from 43 to 1,919 words (mean 457), spanning brief news items to in-depth analyses. This diverse corpus provides a robust basis for evaluating GBV representation across content types.

3.3 Annotation process

To support model training and evaluation, we developed a comprehensive annotation framework based on our taxonomy of 32 expert-defined criteria across 9 categories. We created detailed annotation guidelines with specific examples and edge cases for each criterion. Each of the 32 expert-defined criteria was operationalized as a numeric weighting criterion using a binary scale. This uniform coding scheme ensures consistency across criteria and allows straightforward computation of evaluation metrics such as accuracy and F1-score. [1] provide methodological foundations for assessing inter-coder agreement in computational linguistics, which we applied to ensure reliability in our annotation process.

A team of three subject-matter experts in journalism, gender studies, and media analysis annotated the dataset. For the gold-standard test set (300 randomly selected articles), each article was independently labeled by at least two annotators. Inter-annotator agreement, measured using Cohen's kappa, averaged 0.81, indicating substantial agreement across criteria. For criteria with lower initial agreement ($\kappa < 0.7$), calibration sessions were held to resolve discrepancies and refine the annotation guide. This iterative process improved consistency without oversimplifying expert judgment.

The remaining articles were distributed for single annotation, with 20% overlap to monitor reliability throughout. This dataset serves two functions: as ground truth for evaluating model performance and as a source of in-context examples for few-shot prompting in the hybrid framework.

4 Methodology and implementation

4.1 Expanded evaluation framework

We expanded our previous evaluation framework from 27 to 32 criteria, organized into nine categories to enable more targeted and interpretable analysis of GBV media coverage:

1. Classification and Context of Violence (**CCV**): Identification and contextualization of different forms of violence.
2. Article Structure and Presentation (**ASP**): Headline framing, article layout, and citation practices.
3. Reporting Balance (**RB**) – Narrative symmetry between victims and perpetrators.
4. Societal Context (**SC**): Broader social framing and systemic patterns.
5. Victim Characterization (**VC**): Description of victims, including implicit victim-blaming.
6. Perpetrator Characterization (**PC**): How perpetrators are described or justified.
7. Power Dynamics (**PD**): References to control, coercion, or power imbalances.
8. Psychological and Health Factors (**PHF**): Inclusion and framing of mental health or addiction.
9. Linguistic Choices (**LC**): Language analysis, including euphemisms, passive constructions, and emotional tone.

Each criterion is accompanied by guidelines and examples to ensure consistent coding. For example, Linguistic Choices goes beyond word-level analysis to examine how grammar and metaphors shape narratives, revealing subtle biases.

Table 1 presents the distribution of our criteria across categories, showing actual criteria from our framework.

In our hybrid approach, each evaluation criterion is handled by the method (rule-based NLP or LLM) that performs best for it. For instance, in Societal Context, statistical references are captured by NLP, whereas nuanced cultural

Table 1. Distribution of evaluation criteria by category

Category	Count	Example criteria
CCV	3	Type and context of violence, mention of femicide
ASP	5	Title framing, sourcing, case linkage, accuracy
RB	3	Expert consultation, victim resources
SC	4	Statistical framing, systemic context, cultural references
VC	4	Victim behavior, identity, consequences
PC	5	Perpetrator identity, language, consequences
PD	3	Control relationships, escalation patterns
PHF	3	Mental health, substance use, trauma
LC	2	Minimization vs. explicit framing

references (e.g., rape culture) are better interpreted by an LLM. This allocation leverages the precision of deterministic NLP for explicit content and the contextual understanding of LLMs for implicit narratives, yielding a more nuanced and robust analysis than a single-method approach.

4.2 Technical implementation

We adopted a hybrid system architecture that strategically assigns each criterion to either traditional NLP methods or LLMs, depending on which approach yields better performance for that specific task.

Processing architecture Figure 1 illustrates the full processing architecture of our hybrid evaluation system. The system processing pipeline begins with article ingestion and tokenization. Articles then proceed through two parallel analytical pipelines:

- A rule-based NLP pipeline, built in Python using libraries such as SpaCy, NLTK, and scikit-learn.
- An LLM analysis pipeline, integrating multiple models including GPT-4, Llama 3-12B, Gemma 3-12B, and DeepSeek-R1-14B.

Key components of the architecture include:

- Embedding search and key phrase extraction to identify relevant content segments.
- Dynamic prompt generation, incorporating evaluation feedback for continuous optimization.
- Model-specific processing, using standardized few-shot prompting strategies for LLMs.
- Results integration, applying a weighted ensemble logic—using criterion-specific weights derived from validation accuracy—to consolidate outputs.

This architecture promotes modular scalability and analytical flexibility, allowing future model upgrades or the addition of other language pipelines. As

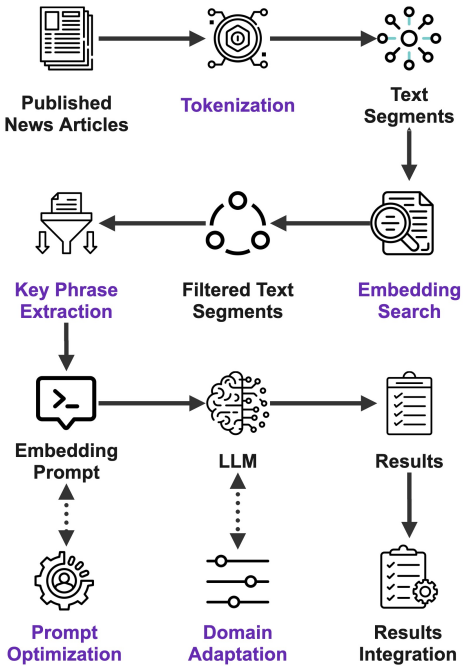

Fig. 1. System architecture integrating NLP and LLM pipelines

shown in Figure 1, the system integrates both rule-based and language model-driven processes, from article ingestion and segmentation to final evaluation synthesis, with feedback loops for continual prompt optimization during development and domain adaptation for evaluation. Domain adaptation here means tailoring prompts and instructions to the task, not fine-tuning model weights.

NLP implementation For the 14 criteria suited to structured and terminological analysis, we used a deterministic NLP approach. Components included:

- French-language lemmatization and stemming.
- Named Entity Recognition with custom tags relevant to GBV.
- TF-IDF and contextual embeddings for statistical analysis.
- Custom sentiment calibration models for sensitive topics.

These techniques proved especially effective in structurally well-defined categories, where traditional NLP methods achieved high accuracy—up to 96% for article-related structure and 99% for linguistic framing—demonstrating the strength of rule-based approaches when patterns are explicit and consistent.

The high performance in these categories can be attributed to the explicit, pattern-based nature of the criteria involved, which align well with the strengths of deterministic NLP techniques. However, as noted in our results, NLP methods

struggled with more interpretive categories such as Reporting Balance, Victim Characterization and Power Dynamics, where implicit language and framing devices required deeper semantic understanding.

LLM implementation and comparison We evaluated each LLM's performance across the 18 remaining criteria, focusing on contextual interpretation, ambiguity resolution, and implicit framing.

- GPT-4 served as the baseline and demonstrated high performance, especially on nuanced categories like Victim and Perpetrator Characterizations.
- Gemma 3-12B showed strong generalization across tasks, particularly in multilingual prompts.
- DeepSeek-R1-14B performed well in inferencing Psychological and Health Factors.
- Llama 3-12B, while more lightweight, delivered competitive results in categories with shorter content structures.

All LLMs were tested using the few-shot prompting strategy proposed by [8], incorporating our own expert-annotated examples and domain-specific context. Prompt engineering was an iterative development process guided by performance logs and expert feedback. The best-performing prompt from this process was then fixed and used for all final evaluations.

Hybrid approach integration Rather than relying on uniform application, we developed a decision framework to assign each criterion to its optimal analytical technique. Criteria were evaluated using performance benchmarks and qualitative error analyses. The results integration layer used:

- Criterion-specific weighting, based on accuracy scores from validation sets.
- Ensemble averaging, where both NLP and LLM outputs contributed proportionally.

We selected accuracy as our primary evaluation metric because our annotation framework was designed to achieve balanced class distributions across evaluation criteria, making accuracy a straightforward and intuitive representation of model correctness. In contrast, metrics like F1-score are typically preferred for imbalanced datasets to balance precision and recall, a scenario less applicable here due to our careful criterion selection and data balancing strategies. Moreover, accuracy provides clearer interpretability, particularly when communicating results to media professionals.

This hybrid configuration outperformed any single-method system. The modular design also facilitates future research extensions, such as integrating multimodal data or adapting criteria for cross-linguistic media studies.

5 Experimental setup and results

5.1 Experimental environment

To ensure a fair and reproducible evaluation, all experiments were conducted in a controlled environment using Python 3.12. NLP components leveraged libraries such as SpaCy 3.5, scikit-learn, and NLTK. Contextual embeddings were generated using the SentenceTransformers library with multilingual models.

LLMs were accessed via standardized APIs or deployed locally using optimized configurations. GPT-4 was accessed through OpenAI's API with temperature set to 0.2 to prioritize deterministic outputs. Open-source models—Llama 3-12B, Gemma 3-12B, and DeepSeek-R1-14B—were run on local inference servers with 16-bit precision and prompt caching for efficiency.

We selected these models to compare performance across both commercial and open-source options. GPT-4 served as a high-performance benchmark. Gemma and Llama were chosen for their strong multilingual capabilities and accessibility, while DeepSeek was included for its reported strength in inferential reasoning.

All models were evaluated on identical inputs, and all code, prompts, and outputs were version-controlled for reproducibility.

5.2 Evaluation methodology

We employed a multi-metric evaluation strategy centered on a manually annotated gold-standard test set. This set consisted of 300 articles randomly sampled from our corpus, each annotated by at least two domain experts using the 32 evaluation criteria across 9 categories. Model selection was not based on this annotated dataset. Instead, we selected GPT-4, Gemma, Llama, and DeepSeek based on prior benchmarks, reported domain strengths, and accessibility considerations. The 300-article set was used exclusively as a held-out test set for final benchmarking, ensuring no overlap between model choice and evaluation data.

Evaluation metrics included:

- Accuracy: Correct classifications over total evaluations.
- Precision and Recall: Measured per criterion, averaged by category.
- F1-score: Harmonic mean of precision and recall.

For LLM-based tasks, we used few-shot prompting with in-context examples drawn from annotated data. NLP components were benchmarked using predefined rules and validated classifiers. To assess the statistical significance of performance differences between methods, we conducted paired t-tests with a significance level of $\alpha = 0.05$.

5.3 Comparative model analysis

We conducted a detailed performance comparison across traditional NLP techniques, GPT-4, and three open-source LLMs (Gemma 3-12B, Llama 3-12B,

and DeepSeek-R1-14B). This evaluation spanned all 32 criteria, yielding both category-level and individual criterion performance metrics. Table 2 presents the core performance results of the hybrid system, which achieved an overall accuracy of 87.1% – an 11 percentage point improvement over our previous framework's 76.1% accuracy. The precision, recall, and F1-score are also high, reflecting the system's robust and balanced performance. The hybrid approach maintained consistently strong results across evaluation dimensions, with particularly high accuracy in certain categories (notably Linguistic Choices and Article Structure and Presentation).

Table 2. Performance metrics for the hybrid system

Metric	Score
Accuracy	87.1%
Precision	83.2%
Recall	79.5%
F1-score	77.7%

Traditional rule-based NLP techniques excelled in structurally well-defined categories. For example, the NLP-only pipeline achieved near-perfect accuracy in Linguistic Choices (99%), Article Structure and Presentation (96%), and Psychological and Health Factors (95%). However, NLP struggled on criteria requiring deeper semantic understanding and context: it performed poorly on Reporting Balance (32% accuracy), Victim Characterization (36%), and Power Dynamics (39%), where implicit language and framing cues are crucial. In contrast, GPT-4 handled these interpretive dimensions much more effectively – for instance, GPT-4 attained 81% accuracy in Victim Characterization, 88% in Perpetrator Characterization, and 83% in Classification and Context of Violence. The open-source LLMs also showed solid (if slightly lower) performance, with overall accuracies of 75.0% for Gemma 3-12B, 72.4% for DeepSeek-R1-14B, and 67.0% for Llama 3-12B. Each model demonstrated unique strengths in certain categories, as summarized in Table 3.

Table 3. Model accuracy and strongest performing categories

Model	Accuracy	Strongest categories
GPT-4	77.9%	CCV, PC
Gemma 3-12B	75.0%	PC, PHF
DeepSeek-R1-14B	72.4%	ASP, VC
Llama 3-12B	67.0%	ASP, LC

No single model or method excelled across all dimensions; each has specific strengths (for example, GPT-4 in nuanced interpretation, Gemma in multilingual context adaptation, and NLP in structure-sensitive patterns). These com-

plementary strengths motivated our hybrid approach. Indeed, the results validate assigning each evaluation criterion to the most suitable method: the rule-based NLP was most effective for explicit, repetitive patterns (e.g. Linguistic Choices and Article Structure), while LLMs outperformed on context-dependent criteria (e.g. Victim Characterization and Reporting Balance) that require implicit bias detection and broader reasoning.

Table 4 provides a detailed accuracy breakdown per category for each method, including the hybrid. As shown, the hybrid model achieves the highest accuracy in every category, effectively combining the advantages of both approaches. Notably, the hybrid reaches 87% accuracy in Reporting Balance and 86% in Victim Characterization, dramatically higher than the NLP-only results for those categories (which were 32% and 36%, respectively). Overall, the hybrid approach consistently matches or exceeds the best individual model in each category.

Table 4. Per-category accuracy comparison across models

Category	**NLP only**	**GPT-4**	**Llama-3**	**Gemma-3**	**DeepSeek-R1**	**Hybrid**
CCV	54%	83%	69%	78%	73%	89%
ASP	96%	81%	75%	72%	80%	95%
RB	32%	82%	54%	65%	72%	87%
SC	74%	65%	68%	75%	66%	79%
VC	36%	81%	62%	77%	77%	86%
PC	63%	88%	66%	82%	70%	88%
PD	39%	78%	67%	76%	68%	77%
PHF	95%	68%	70%	78%	78%	86%
LC	99%	75%	72%	72%	68%	99%
Overall average	**65%**	**78%**	**67%**	**75%**	**72%**	**87%**

Statistical tests (paired t-tests, $\alpha = 0.05$) confirm that the hybrid model's improvement is significant. The hybrid's overall accuracy advantage over every individual method is statistically significant ($p < 0.01$). It also achieved uniformly high accuracy across all categories, with stand-out performance in Linguistic Choices (99%), Article Structure and Presentation (95%), and Psychological and Health Factors (86%). The only category where the rule-based NLP was on par with the hybrid is Linguistic Choices (both 99% accuracy, difference not significant). In addition, GPT-4 and DeepSeek each came close to the hybrid on one category (Article Structure and Societal Context, respectively), but those differences were not statistically significant. These findings underscore the benefit of selectively combining rule-based and neural techniques to leverage their complementary strengths in media analysis.

Compared to a purely LLM-driven approach, our hybrid system shows clear advantages. If an LLM alone were used for all criteria, performance would suffer in areas where structured analysis is crucial. For example, GPT-4 alone achieved about 78% overall accuracy (see Table 4), noticeably below the hybrid's 87%. It also struggled with structurally defined categories – attaining only 81% in

Article Structure (versus 95% by the hybrid). These gaps illustrate that even a state-of-the-art LLM cannot fully replicate the hybrid's balanced strengths across diverse criteria.

Furthermore, to our knowledge no existing hybrid system in the literature directly mirrors our integration of rule-based NLP with LLMs for media analysis, underlining the novelty of this framework. One could alternatively employ a transformer model fine-tuned specifically for fairness or media framing tasks. While such a specialized model (e.g., a BERT variant trained on bias detection) can excel at targeted classifications, it would address a narrower set of biases or frames and require extensive labeled data for each nuance. In contrast, our hybrid method leverages a general-purpose LLM's broad knowledge alongside rule-based precision, covering a wide spectrum of 32 criteria without task-specific retraining. This comparative perspective confirms that our combined approach captures nuanced media representations more comprehensively than either a purely generative LLM pipeline or a single fine-tuned model alone.

6 Discussion

6.1 Interpretation of results

The findings confirm that a hybrid evaluation system, combining NLP and LLMs, can approach human-level performance in assessing GBV representation in media. Traditional NLP techniques remain highly effective in handling structurally explicit criteria, especially when patterns are rule-based or linguistically predictable. However, their limitations become apparent when faced with subtle semantic framing, implicit bias, or nuanced social commentary—areas where LLMs, particularly GPT-4, excel.

The comparative analysis demonstrated the added value of open-source LLMs, with Gemma 3-12B and DeepSeek-R1-14B providing competitive alternatives to GPT-4. Their performance in select categories—such as Psychological and Health Factors and Societal Context—suggests that with appropriate prompting and domain-specific context, open models can offer cost-effective and privacy-friendly options for large-scale media analysis.

Our results also highlight the importance of tailored evaluation strategies. By assigning criteria based on the analytical strengths of each method and refining prompts and embeddings accordingly, we achieved substantial performance gains (+11 points over baseline). This validates the hypothesis that hybrid integration is not only feasible but advantageous in socially sensitive AI applications.

6.2 Practical implications

This study contributes to the emerging field of algorithmic media evaluation by offering a scalable, adaptable framework for assessing gender representation in journalism. Newsrooms, NGOs, and media regulators could adopt this framework to monitor coverage of GBV more systematically. Because the system is designed

with modular inputs, it can be adapted to other languages, countries, or thematic focuses with limited reconfiguration.

From a technological standpoint, our approach underscores the potential for synergistic use of deterministic and generative models. As LLMs continue to evolve and become more accessible, integrating them strategically with rule-based systems will offer flexible solutions in both high- and low-resource environments.

6.3 Limitations and future work

Several limitations warrant attention. First, although our dataset is large and diverse, it remains geographically bounded to Swiss French-language media, limiting generalizability. Second, while expert annotations provide a solid ground truth, some evaluation criteria remain interpretative, potentially introducing variability in LLM outputs. Third, the system struggled with certain content types: short articles (under 150 words) often lacked sufficient context for reliable assessment, and dialectal variations (regionalisms in Swiss French) caused vocabulary mismatches that undermined NLP accuracy. Fourth, hybrid narrative formats blending reporting with opinion introduced additional ambiguity, particularly in evaluating criteria like balance and systemic framing; these challenges persisted despite targeted mitigation efforts (e.g., length-sensitive thresholds and regional lexicons) that achieved only moderate success, highlighting the complexity of automating nuanced media evaluations.

Future research could address these limitations by:

- Extending the framework to multilingual and cross-regional media corpora.
- Incorporating human-in-the-loop evaluation to combine expert feedback with automated analysis.
- Enhancing interpretability tools for LLM outputs.
- Exploring multimodal media analysis by integrating image and video content.

7 Conclusion

This paper presented a novel hybrid AI framework for evaluating gender representation in media reporting on gender-based violence. By combining NLP and LLM methods across 32 expert-defined criteria in 9 categories, our system achieves a new state-of-the-art in this domain. The approach is not only more accurate than prior methods but also more flexible and extensible.

Beyond performance metrics, this framework demonstrates how different computational approaches—deterministic and generative—can be strategically integrated to complement each other. Such a methodology is especially valuable in analyzing socially sensitive topics, where linguistic nuance and contextual framing are essential.

Our results suggest not only the feasibility of hybrid automation for media analysis but also its relevance for promoting more equitable journalism practices. As AI continues to evolve, interdisciplinary solutions like ours will be central to ensuring that technological tools contribute meaningfully to democratic discourse, ethical media practices, and public accountability.

Acknowledgments. This research was supported by the University of Applied Sciences and Arts Western Switzerland (HES-SO) under grant number 133851. We are especially grateful to our project partner DécadréE, and in particular to its director, Valérie Vuille, for their dedicated collaboration and invaluable support.

References

1. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational linguistics **34**(4), 555–596 (2008)
2. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 610–623 (2021)
3. Berns, N.: Degendering the problem and gendering the blame: Political discourse on women and violence. Gender & society **15**(2), 262–281 (2001)
4. Byerly, C.M., Ross, K.: Women and media: A critical introduction. John Wiley & Sons (2008)
5. Carll, E.K.: News portrayal of violence and women: Implications for public policy. American Behavioral Scientist **46**(12), 1601–1610 (2003)
6. Chen, Y., Zhai, Y., Sun, S.: The gendered lens of ai: examining news imagery across digital spaces. Journal of Computer-Mediated Communication **29**(1), zmad047 (2024)
7. COE: The council of europe convention on preventing and combating violence against women and domestic violence. (2011), https://www.coe.int/en/web/gender-matters
8. Dang, H., Mecke, L., Lehmann, F., Goller, S., Buschek, D.: How to prompt? opportunities and challenges of zero-and few-shot learning for human-ai interaction in creative applications of generative models. arXiv preprint arXiv:2209.01390 (2022)
9. DécadréE: Report 2023: Media coverage of sexist violence. (2023), https://decadree.com/wp-content/uploads/2023/11/2023_Rapport_ViolenceSexistes_NV-1.pdf
10. Easteal, P., Holland, K., Judd, K.: Enduring themes and silences in media portrayals of violence against women. In: Women's Studies International Forum. vol. 48, pp. 103–113. Elsevier (2015)
11. Flood, M., Pease, B.: Factors influencing attitudes to violence against women. Trauma, violence, & abuse **10**(2), 125–142 (2009)
12. Forja-Pena, T., García-Orosa, B., López-García, X.: A shift amid the transition: Towards smarter, more resilient digital journalism in the age of ai and disinformation. Social Sciences **13**(8), 403 (2024)
13. Liu, Z., Darbellay, A., Balet, N.G., Vuille, V.: Advancing gender equality in media: Tackling stereotypes and biases with ai. In: International KES Conference on Intelligent Decision Technologies. pp. 413–424. Springer (2024)

14. Manasi, A., Panchanadeswaran, S., Sours, E.: Addressing gender bias to achieve ethical ai. (2023), https://theglobalobservatory.org/2023/03/gender-bias-ethical-artificial-intelligence/
15. O'Connor, S., Liu, H.: Gender bias perpetuation and mitigation in ai technologies: challenges and opportunities. AI & SOCIETY **39**(4), 2045–2057 (2024)
16. Ross Arguedas, A., Mukherjee, M., Kleis Nielsen, R.: Women and leadership in the news media 2024: Evidence from 12 markets. (2024), https://reutersinstitute.politics.ox.ac.uk/women-and-leadership-news-media-2024-evidence-12-markets
17. Simon, F.: Artificial intelligence in the news: How ai retools, rationalizes, and reshapes journalism and the public arena (2024)
18. Tilawat, M.: Ai bias report 2025: Llm discrimination is worse than you think! (2025), https://www.allaboutai.com/resources/ai-statistics/ai-bias/
19. WHO: Violence against women. (2024), https://www.who.int/news-room/fact-sheets/detail/violence-against-women