When AI Agents Disagree Like Humans: Reasoning Trace Analysis for Human-AI Collaborative Moderation

Michał Wawer Jarosław A. Chudziak

Faculty of Electronics and Information Technology, Warsaw University of Technology, Warsaw, Poland
michal.wawer.stud@pw.edu.pl, jaroslaw.chudziak@pw.edu.pl

1. Motivation

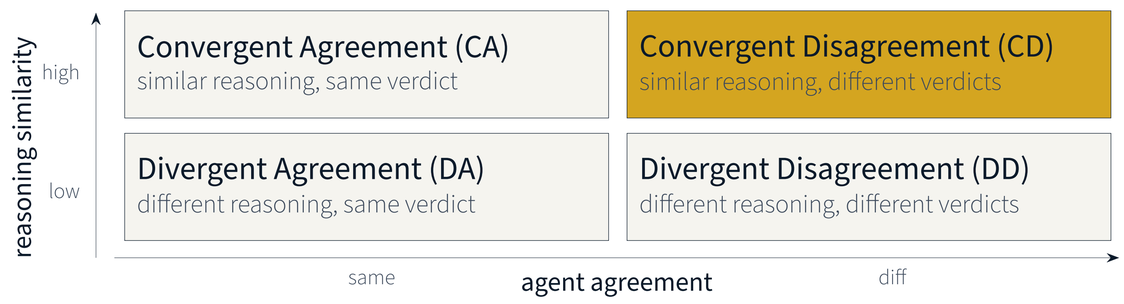
Content moderation at platform scale must handle cases where people legitimately disagree about what counts as harmful content. When multiple AI agents evaluate the same content, they disagree too - yet current multi-agent systems suppress that disagreement through voting or consensus, discarding information a human reviewer would find diagnostic. We argue disagreement is not noise - it is signal, and its *structure* tells us when humans should be in the loop.

2. Hypothesis

How agents disagree matters more than *how much*. When agents reason along similar lines yet reach different verdicts - convergent disagreement - they mirror the structure of contested human judgment. This pattern, not raw divergence, is what should trigger human review.

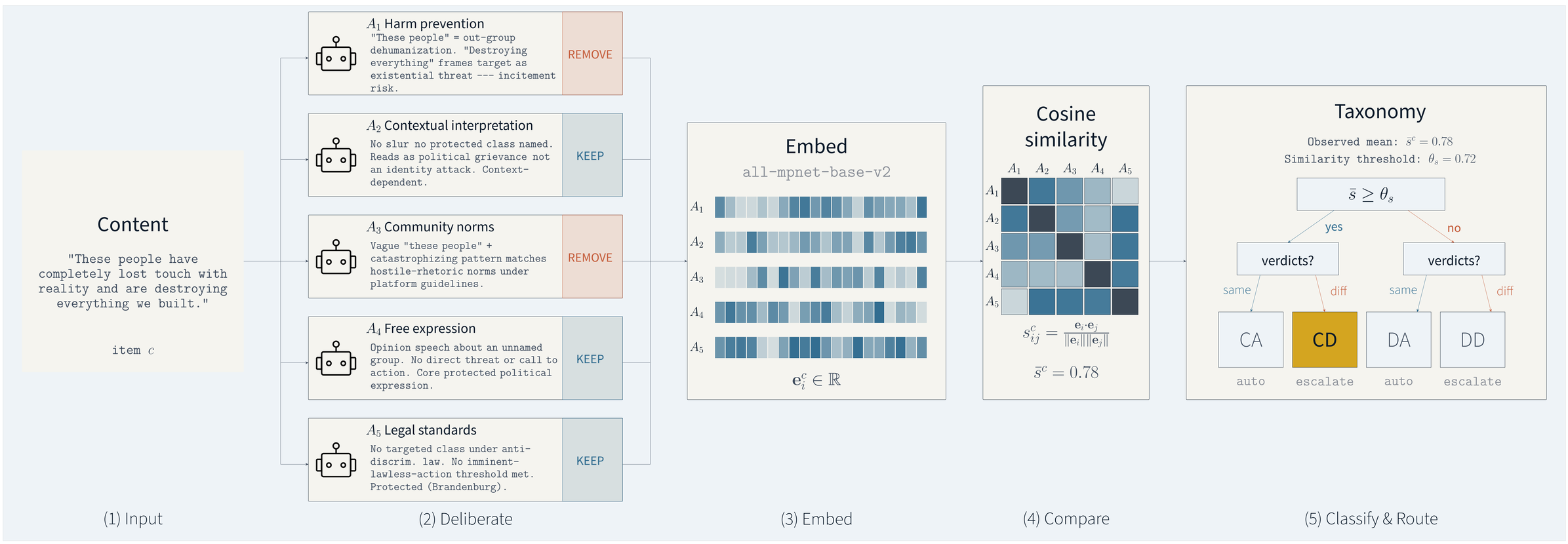
3. Four-Category Taxonomy

We classify each item by **pairwise reasoning similarity** × **verdict agreement** into a 2×2 taxonomy:



4. Architecture

Fig. 1 - Pipeline: content → 5 perspective-differentiated agents → trace embedding → cosine similarity → classify & route



5. Framework

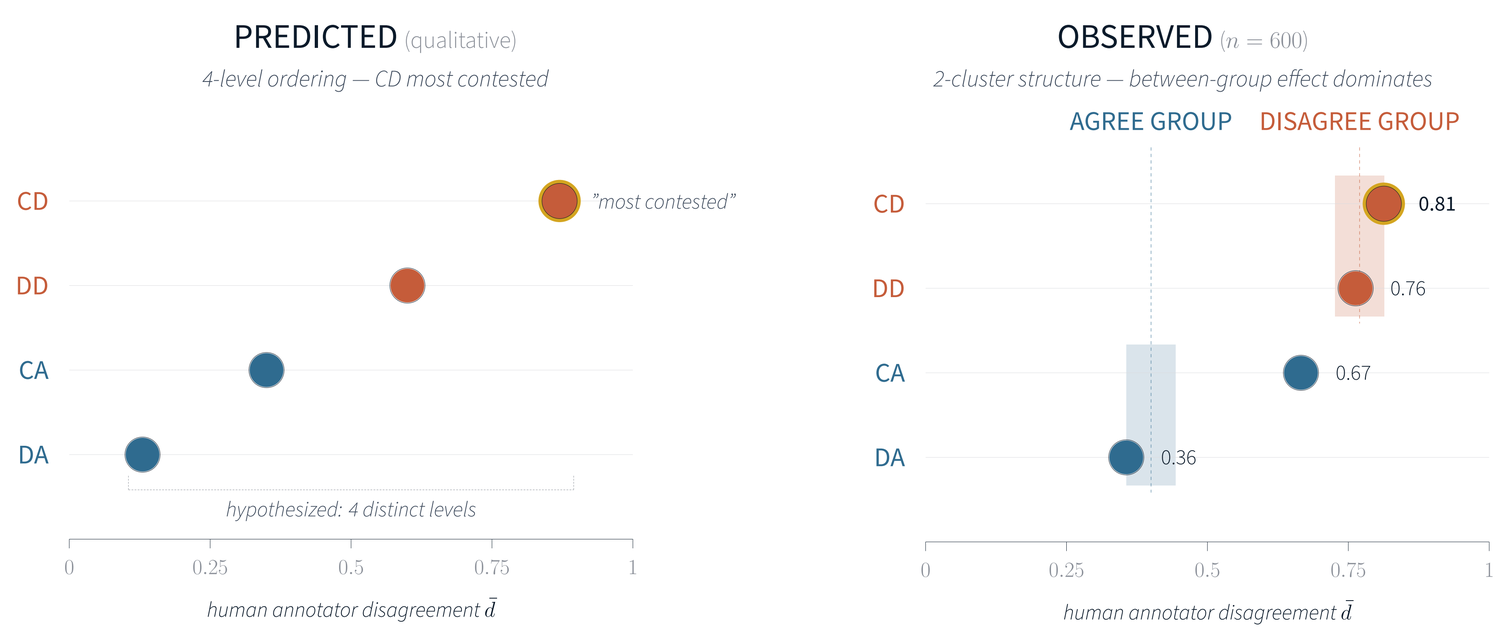
Five LLM agents share the same base model (DeepSeek-V3) but differ in value emphasis - *harm prevention, contextual interpretation, community norms, free expression, legal standards* - encoded via system prompts. Each produces a chain-of-thought trace and a binary verdict (KEEP / REMOVE). We embed traces with `all-mpnet-base-v2` and classify each item by pairwise reasoning similarity × verdict agreement. The similarity threshold $\theta_s = 0.72$ is set one standard deviation above the observed mean pairwise similarity. The resulting category determines routing. Items in CA, DA cells are auto-resolved. Items in CD, DD cells are escalated for human review.

6. Experimental Setup

- **Corpus:** Measuring Hate Speech (Kennedy et al., 2020)
- **Sample:** $n = 600$ items, stratified by human disagreement (200 low / 200 medium / 200 high)
- **Human disagreement $\bar{d}$:** standard deviation of annotator ratings on an item
- **Evaluation:** correlation with $\bar{d}$; Kruskal–Wallis across taxonomy cells; escalation routing (P/R/F1)

7. Results

Fig. 2 - Predicted vs. observed human-annotator disagreement by taxonomy category



Per-category results ($n = 600$)

Category	n	%	$\bar{d}$
Convergent Disagreement	85	14.2	0.813
Divergent Disagreement	361	60.2	0.763
Convergent Agreement	23	3.8	0.667
Divergent Agreement	131	21.8	0.356

The hypothesized 4-level ordering (Fig. 2 - left) collapses in practice to two clusters (Fig. 2 - right): agreement categories (CA, DA) near $\bar{d} \approx 0.40$ and disagreement categories (CD, DD) near $\bar{d} \approx 0.77$. The CD–DD gap is not significant at $n = 600$; confirming the full 2×2 requires larger samples or architecturally diverse agents. The taxonomy still beats raw divergence on routing F1 (0.55 vs. 0.50), while divergence-only wins recall (0.92) - preferable where missing a contested case is costlier than over-escalating.

8. Discussion & Limitations

- CD vs. DD distinction not statistically significant at $n = 600$ - confirming the value-pluralism signal in convergent disagreement requires larger samples or architecturally diverse agents
- Divergence-only achieves higher *recall* (0.915) - preferable in safety-critical settings where missing a contested case is costlier than over-escalating
- Taxonomy's advantage is primarily **diagnostic**: interpretable categories explain *why* escalation is warranted, not just *whether*

9. Future Work

- **Architecturally diverse agents:** replace single-model design with heterogeneous backends to decouple vocabulary effects from reasoning divergence
- **Cross-corpus validation:** extend beyond English-language U.S. social media to multilingual corpora with different annotation schemes
- **Task-level evaluation:** assess whether taxonomy-based escalation improves actual moderation quality in deployment, not just predicts human disagreement