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,jaroslaw.chudziak}@pw.edu.pl

ABSTRACT

When LLM-based multi-agent systems disagree, current practice treats this as noise to be resolved through consensus. We propose it can be signal. We focus on hate speech moderation, a domain where judgments depend on cultural context and individual value weightings, producing high legitimate disagreement among human annotators. We hypothesize that convergent disagreement, where agents reason similarly but conclude differently, indicates genuine value pluralism that humans also struggle to resolve. Using the Measuring Hate Speech corpus, we embed reasoning traces from five perspective-differentiated agents and classify disagreement patterns using a four-category taxonomy based on reasoning similarity and conclusion agreement. We find that raw reasoning divergence weakly predicts human annotator conflict, but the structure of agent discord carries additional signal: cases where agents agree on a verdict show markedly lower human disagreement than cases where they do not, with large effect sizes ($d > 0.8$) surviving correction for multiple comparisons. Our taxonomy-based ordering correlates with human disagreement patterns. These preliminary findings motivate a shift from consensus-seeking to uncertainty-surfacing multi-agent design, where disagreement structure - not magnitude - guides when human judgment is needed.

1 INTRODUCTION

Content moderation presents a fundamental challenge for automated systems: decisions must be made at massive scale while navigating genuine disagreement about what constitutes harmful content (Gorwa et al., 2020; Kuo et al., 2023). A comment referencing cultural practices may appear offensive to moderators unfamiliar with the context but appropriate to those within the community. These are not edge cases to be engineered away but endemic features of the moderation task (Hartmann et al., 2025; Tomasev et al., 2021; Rieder & Skop, 2021). Multi-agent architectures have emerged as a promising approach to this challenge (Harbar & Chudziak, 2025; Zamojska & Chudziak, 2025), leveraging collective reasoning among multiple AI agents to improve decision quality (Du et al., 2024; Chen et al., 2024b). Current implementations treat inter-agent disagreement primarily as a problem to overcome through consensus protocols (Zheng et al., 2025), weighted voting (Jo & Park, 2025; Luo et al., 2025) or Byzantine fault tolerance mechanisms that filter disagreeing agents as potentially unreliable (Chen et al., 2024a). We argue this approach discards valuable information. When AI agents with similar capabilities persistently disagree about a judgment, this disagreement itself carries diagnostic meaning: it may signal that the content occupies a contested semantic region where human reasoners would also disagree. Rather than treating such disagreement as noise to be eliminated, we propose that AI disagreement can serve as a proxy for human cognitive disagreement, enabling systems to recognize when to defer to human judgment.

We connect collective intelligence research to multi-agent AI design, identifying conditions under which agent disagreement mirrors patterns of human cognitive disagreement. Our central hypothesis is that disagreement patterns in multi-agent deliberation can be systematically analyzed to distinguish cases requiring automated resolution from those requiring human judgment, transforming multi-agent moderation from a consensus-seeking system into an uncertainty-surfacing one.

Based on it, we introduce a methodology for analyzing reasoning trace divergence that distinguishes value-based disagreement from error-based disagreement, paralleling how humans distinguish principled disagreement from mere confusion. Our initial results reveal a counterintuitive pattern: the magnitude of reasoning divergence between agents correlates negatively with human disagreement ($r = -0.19$), suggesting that how much agents diverge is uninformative. However, *how* they diverge matters, our four-category taxonomy separates cases by human disagreement level (Kruskal-Wallis $H = 54.0$), with the agreement-disagreement boundary driving the effect. This motivates a shift from measuring divergence to characterizing its structure.

2 THEORETICAL FRAMEWORK

Human disagreement about evaluative judgments often reflects genuine value pluralism rather than error (Gajewska et al., 2026; Wang & Zhu, 2022). Cognitive science research demonstrates that humans weigh competing values differently when reasoning about complex social situations (Liao, 2011; Sadowski & Chudziak, 2025). Whether strong political speech constitutes harassment depends on how one weighs free expression against community protection and different people applying consistent principles may disagree. In AI systems outputs, rather than assuming a ground truth that disagreeing agents are failing to reach consensus, we can treat certain disagreement patterns as informative signals that content involves genuine value tensions (Wang et al., 2025b;a).

Recent work confirms these concerns. Studies of multi-agent debate find that majority pressure can suppress independent correction (Huang et al., 2025), with extended debate rounds sometimes entrenching errors rather than correcting them (Liang et al., 2024). Evaluations across multiple debate frameworks find they "fail to consistently outperform simpler single-agent strategies" (Zhang et al., 2025) suggesting multi-agent deliberation may degrade to inefficient resampling without genuine epistemic diversity (Du et al., 2024).

We propose distinguishing two sources of persistent disagreement, inspired by cognitive accounts of human reasoning: (1) Value pluralism (convergent disagreement): Agents reasoning from different implicit value weightings reach different conclusions through valid reasoning chains. When agents understand content similarly but judge it differently, the disagreement reflects legitimate value tensions, mirroring how humans with shared understanding can reach different evaluative conclusions. (2) Noise or error (divergent disagreement): Agents make inconsistent judgments due to different interpretations, context gaps or reasoning failures (Kostka & Chudziak, 2025). This parallels cases where human disagreement stems from miscommunication rather than genuine value differences. The methodological challenge is distinguishing these sources. We propose that analyzing the structure of disagreement, not just its presence, enables better classification for content moderation.

Our approach connects to formal results in social choice and collective intelligence. The Condorcet Jury Theorem (Austen-Smith & Banks, 1996) predicts that majority voting among independent agents converges to correct decisions, but recent work shows LLM agents violate the independence assumption due to shared pretraining, causing consensus without correctness gains (Denisov-Blanch et al., 2025). This failure mode motivates our shift from consensus-seeking to disagreement analysis. More broadly, judgment aggregation theory (List & Pettit, 2002) demonstrates that aggregating premise-level and conclusion-level votes can yield contradictory collective judgments, the discursive dilemma. Our taxonomy operationalizes a related insight: agents may agree on premises (high reasoning similarity) yet disagree on conclusions, and this structural pattern carries diagnostic value that raw aggregation discards.

3 METHODOLOGY

We employ a deliberation architecture with $N = 5$ agents instantiated from the same base model (DeepSeek-V3) but differentiated through system prompts encoding distinct moderator perspectives. Each perspective emphasizes different values: harm prevention, contextual interpretation, community norms, free expression, and legal standards, respectively.

Design shown on Fig. 1 creates simulated perspective diversity while controlling for capability differences, all agents share the same underlying knowledge and reasoning capacity, differing only in value emphasis and framing. The distinctive vocabularies enable semantic differentiation of rea-

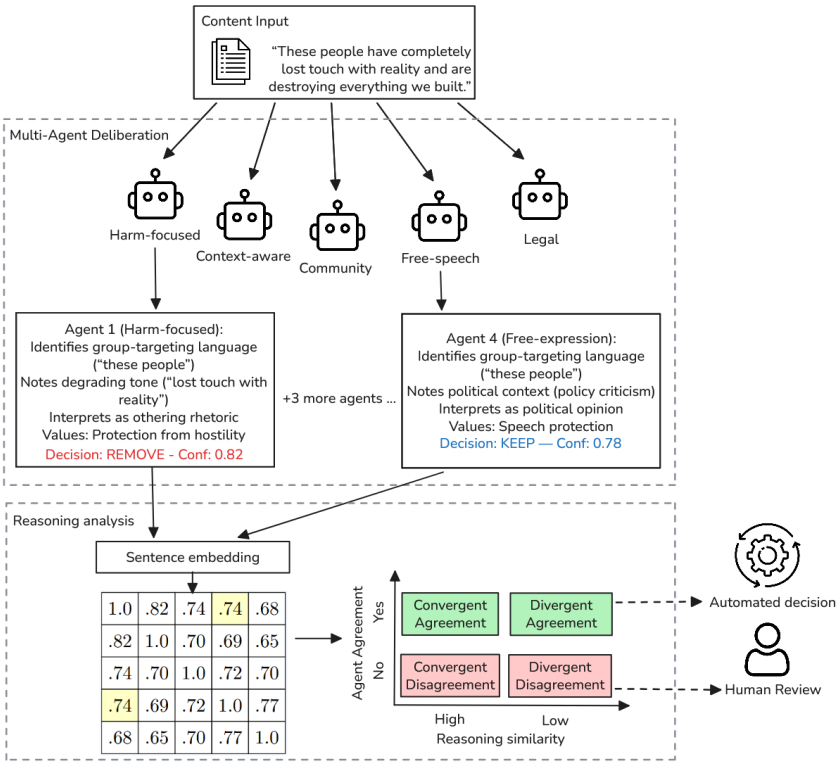

Figure 1: Overview of our framework design, content is processed by N perspective-differentiated agents, each generating reasoning traces. Traces are embedded and compared via cosine similarity. The disagreement taxonomy routes cases to automated resolution or human escalation.

soning traces. For each content item, agents independently generate moderation decisions with explicit reasoning traces. We invoke structured justifications through chain-of-thought prompting (Wei et al., 2022), requiring agents to articulate: (1) content interpretation, (2) relevant considerations, (3) value trade-offs, and (4) final judgment with confidence. Agents produce binary decisions (REMOVE/KEEP) with confidence scores and explicit statements of which values drove their judgment.

Reasoning traces are the intermediate justifications agents generate when arriving at decisions. Our methodology for analyzing these traces builds on Tajik et al. (2026), who introduced reasoning trace analytics with categories analogous to ours (within-align/within-misalign). We extend their framework to content moderation, where we interpret convergent disagreement as a signal of value pluralism rather than a generic misalignment indicator. We embed traces into a shared semantic space for quantitative comparison. For each agent i and content item c, let r_i^c denote the full reasoning trace. We compute sentence embeddings using a pre-trained transformer model (all-mpnet-base-v2), yielding vector representations $\mathbf{e}_i^c \in \mathbb{R}^d$. We then compute pairwise cosine similarity:

$$s_{ij}^c = \frac{\mathbf{e}_i^c \cdot \mathbf{e}_j^c}{\|\mathbf{e}_i^c\| \|\mathbf{e}_j^c\|} \tag{1}$$

From pairwise similarities, we derive aggregate divergence metrics for each content item:

- Mean similarity: $\bar{s}^c = \frac{2}{N(N-1)} \sum_{i<j} s_{ij}^c$
- Minimum similarity: $s_{\min}^c = \min_{i<j} s_{ij}^c$

Low mean similarity indicates agents are reasoning about the content in fundamentally different ways; low minimum similarity indicates at least one agent pair has highly divergent reasoning. We

classify each content item into one of four categories based on the joint distribution of reasoning similarity and conclusion agreement:

1. Convergent Agreement (high similarity, same conclusion): Agents reason similarly and agree, representing confident decisions where collective intelligence operates normally.
2. Divergent Agreement (low similarity, same conclusion): Agents reason differently but converge on the same judgment through independent reasoning paths.
3. Convergent Disagreement (high similarity, different conclusions): Agents reason similarly but reach different judgments, we hypothesize this is strong signal of value pluralism.
4. Divergent Disagreement (low similarity, different conclusions): Agents reason and conclude differently, representing genuine edge cases or noise from inconsistent interpretation.

4 Experimental Design

We use the Measuring Hate Speech corpus (Kennedy et al., 2020; Sachdeva et al., 2022), which aggregates annotations rating 39,565 social media comments. Annotators rated each comment on 10 ordinal dimensions, using 5-point scales. We compute human annotator disagreement for each content item as the standard deviation of ratings across annotators. Items with high standard deviation represent cases where human perspectives genuinely diverge; items with low standard deviation represent cases of relative consensus. As an initial validation of the proposed framework, we sample $n = 600$ content items stratified by human disagreement level, with equal representation across three bins (200 low, 200 medium, 200 high disagreement). For each item, we:

1. Generate moderation decisions from 5 perspective-differentiated agents via DeepSeek API (temperature $= 0.7$), collecting full reasoning traces
2. Compute pairwise reasoning trace similarity using embeddings (all-mpnet-base-v2)
3. Calculate aggregate divergence as $1 - \bar{s}^c$ (mean pairwise dissimilarity)
4. Classify each item into the four-category taxonomy using similarity threshold $\theta_s = 0.72$ (set at approximately one standard deviation above the observed mean pairwise similarity of 0.67) and majority agreement threshold $> 50\%$
5. Record agent conclusions (binary REMOVE/KEEP decisions)

We evaluate if raw agent reasoning divergence predicts human disagreement by computing Pearson and Spearman correlations between divergence scores and human annotator rating standard deviation. We then examine whether the structure of disagreement carries additional signal by comparing mean human disagreement across the four taxonomy categories and testing between-category differences with Kruskal-Wallis tests. We assess the practical value of the taxonomy for routing decisions by treating high human disagreement as ground truth and comparing category-based escalation against divergence-only and random baselines using precision, recall, and F1, where a true positive occurs when a predictor correctly flags a case that human annotators found genuinely contested.

5 Preliminary Results

Table 1 summarizes the distribution of items across our four-category taxonomy and the mean human annotator disagreement within each category.

Category	**n**	**%**	**Mean** d
Convergent Disagreement (CD)	85	14.2	0.813
Divergent Disagreement (DD)	361	60.2	0.763
Convergent Agreement (CA)	23	3.8	0.666
Divergent Agreement (DA)	131	21.8	0.356

Table 1: Human disagreement (d) by agent disagreement category.

Agent reasoning divergence shows a weak negative correlation with human annotator disagreement (Pearson $r = -0.19$). The overall category effect is significant (Kruskal-Wallis $H = 54.0$). Post-hoc pairwise comparisons reveal that this effect is primarily driven by divergent agreement (DA)

cases showing significantly lower human disagreement than both convergent disagreement (Cohen's $d = 0.89$) and divergent disagreement ($d = 0.80$). The difference between CD ($\bar{d} = 0.813$) and DD ($\bar{d} = 0.763$) is not statistically significant, indicating that the primary diagnostic value of the taxonomy lies in distinguishing agreement from disagreement categories rather than differentiating within disagreement types. We tested sensitivity to θ_s across $[0.60, 0.84]$: the Spearman correlation remains significant at all thresholds ($\rho \in [0.25, 0.30]$ in all cases), confirming that results are robust to the threshold choice.

Predictor	**Precision**	**Recall**	**F1**
Category-based escalation	0.401	0.845	0.548
Divergence Only	0.347	0.915	0.503
Random Baseline	0.333	0.505	0.401

Table 2: Predictor comparison for flagging high human disagreement cases. The escalation score achieves the highest F1; all informed methods outperform random assignment.

The category-based escalation score achieves the highest F1 (0.54), compared to the divergence-only baseline (0.50) and random assignment (0.40). The improvement over divergence-only is modest and may not be significant at this sample size. Notably, the divergence-only predictor achieves higher recall (0.915 vs. 0.845). In safety-critical deployment, where missing a genuinely contested case is costlier than over-escalating, higher recall may be preferable. The taxonomy's advantage is therefore primarily *diagnostic*, it provides interpretable categories explaining *why* escalation is warranted, rather than purely predictive.

6 Discussion and Future Work

Our results suggest that multi-agent systems can implicitly model aspects of human cognitive disagreement, not through the magnitude of their divergence, but through its structure. The weak negative correlation of raw divergence ($r = -0.19$) compared with the positive taxonomy-based correlation ($\rho = 0.27$, escalation F1 $= 0.54$) provides preliminary evidence that *how* agents disagree carries more diagnostic value than *how much*. Both effects are modest, and the CD-DD distinction does not reach statistical significance at the current sample size. The significant finding is the separation between agreement categories (DA, CA) and disagreement categories (DD, CD). The four-category taxonomy currently reduces in practice to a binary: agent agreement vs. disagreement. The full 2×2 structure remains theoretically motivated-convergent disagreement is where value pluralism should be most visible, but confirming this requires larger samples or architecturally diverse agents. The practical implication is a shift in design goals from optimizing for consensus toward optimizing for appropriate routing, where agent disagreement triggers human escalation rather than being suppressed. Our evaluation on 600 items from a single corpus establishes that structural analysis of disagreement shows promise, but substantial further validation is needed. The imbalanced category distribution - DD comprising 60.2% of items, likely reflects our single-model design: because all agents share the same base model and differ only in system prompts, the perspective-specific vocabulary creates systematic semantic divergence even when agents reason substantively similarly. Architecturally diverse agents would likely produce a more balanced distribution by decoupling vocabulary from reasoning effects. Additional limitations include the focus on English-language U.S. social media and the absence of task-level evaluation. We plan to address both through expanded evaluation across multiple corpora and a study assessing whether taxonomy-based escalation improves moderation quality in practice.

7 Conclusion

We proposed treating multi-agent disagreement as diagnostic signal rather than noise in content moderation. By analyzing reasoning trace divergence through a four-category taxonomy, we can distinguish cases likely reflecting genuine value pluralism from those reflecting error — and route them accordingly. This reframes multi-agent moderation from consensus-seeking to collaborative sense-making, where disagreement structure guides the human-AI division of cognitive labor.

REFERENCES

David Austen-Smith and Jeffrey S. Banks. Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review*, 90(1):34–45, 1996.

Bei Chen, Gaolei Li, Xi Lin, Zheng Wang, and Jianhua Li. Blockagents: Towards byzantine-robust llm-based multi-agent coordination via blockchain. In *Proceedings of the ACM Turing Award Celebration Conference - China 2024*, ACM-TURC '24, pp. 187–192, New York, NY, USA, 2024a. Association for Computing Machinery. doi: 10.1145/3674399.3674445.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085, August 2024b. URL https://aclanthology.org/2024.acl-long.381/.

Yuri Denisov-Blanch, Joshua Kazdan, Julia Chudnovsky, Rylan Schaeffer, Shirley Guan, Samuel Adeshina, and Sanmi Koyejo. Consensus is not verification: Why crowd wisdom strategies fail for LLM truthfulness. *arXiv preprint arXiv:2603.06612*, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, 2024.

Ewelina Gajewska, Arda Derbent, Jarosław A. Chudziak, and Katarzyna Budzynska. Algorithmic fairness in nlp: Persona-infused llms for human-centric hate speech detection. In *Proceedings of the 59th Hawaii International Conference on System Sciences (HICSS-59)*, Maui, United States, January 2026. Accepted.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1): 2053951719897945, 2020.

Yarolsav Harbar and Jarosław A Chudziak. Simulating oxford-style debates with llm-based multi-agent systems. In *Asian Conference on Intelligent Information and Database Systems*, pp. 286–300. Springer, 2025.

David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. Lost in moderation: How commercial content moderation apis over-and under-moderate group-targeted hate speech and linguistic variations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–26, 2025.

Jen-Tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Michael Lyu, and Maarten Sap. On the resilience of LLM-based multi-agent collaboration with faulty agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26202–26226. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25ay.html.

Yongrae Jo and Chanik Park. Byzantine-robust decentralized coordination of llm agents, 2025. URL https://arxiv.org/abs/2507.14928.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*, 2020.

Adam Kostka and Jarosław A. Chudziak. Synergizing logical reasoning, knowledge management and collaboration in multi-agent llm system, 2025. URL https://arxiv.org/abs/2507.02170.

Tina Kuo, Alicia Hernani, and Jens Grossklags. The unsung heroes of facebook groups moderation: A case study of moderation practices and tools. 7(CSCW1), April 2023. doi: 10.1145/3579530. URL https://doi.org/10.1145/3579530.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992.

S. Matthew Liao. *Bias and Reasoning: Haidt's Theory of Moral Judgment*, pp. 108–127. Palgrave Macmillan UK, London, 2011. ISBN 978-0-230-30588-5. doi: 10.1057/9780230305885_7. URL https://doi.org/10.1057/9780230305885_7.

Christian List and Philip Pettit. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18(1):89–110, 2002.

Haoxiang Luo, Gang Sun, and Hongfang Yu. Wbft-mllmn: A weighted bft consensus driven multiple large language models network. In *2025 IEEE 2nd International Conference on Deep Learning and Computer Vision (DLCV)*, pp. 1–7, 2025. doi: 10.1109/DLCV65218.2025.11088676.

Bernhard Rieder and Yarden Skop. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of perspective api. *Big Data & Society*, 8(2): 20539517211046181, 2021.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, June 2022. URL https://aclanthology.org/2022.nlperspectives-1.11/.

Albert Sadowski and Jaroslaw A. Chudziak. On verifiable legal reasoning: A multi-agent framework with formalized knowledge representations. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pp. 2535–2545, New York, NY, USA, 2025. Association for Computing Machinery. URL https://doi.org/10.1145/3746252.3761057.

Elham Tajik, Conrad Borchers, Bahar Shahrokhian, Sebastian Simon, Ali Keramati, Sonika Pal, and Sreecharan Sankaranarayanan. Disagreement as data: Reasoning trace analytics in multi-agent systems. In *Proceedings of LAK 2026*, 2026.

Nenad Tomasev, Kevin R McKee, Jackie Kay, and Shakir Mohamed. Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 254–265, New York, NY, USA, 2021. Association for Computing Machinery.

Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=h0ZfDIrj7T.

Leijie Wang and Haiyi Zhu. How are ml-based online content moderation systems actually used? studying community size, local activity, and disparate treatment. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 824–838, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533147. URL https://doi.org/10.1145/3531146.3533147.

Yuxin Wang, Xiaomeng Zhu, Weimin Lyu, Saeed Hassanpour, and Soroush Vosoughi. Impscore: A learnable metric for quantifying the implicitness level of sentences. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=gYWqxXE5RJ.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc., 2022.

Monika Zamojska and Jaros_aw A Chudziak. Games agents play: Towards transactional analysis in llm-based multi-agent systems. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, pp. 1598–1605, 2025.

Hangfan Zhang, Zhiyao Cui, Qiaosheng Zhang, and Shuyue Hu. Multi-LLM-agents debate - performance, efficiency, and scaling challenges. In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL https://openreview.net/forum?id=Wv0J0bEly5.

Lifan Zheng, Jiawei Chen, Qinghong Yin, Jingyuan Zhang, Xinyi Zeng, and Yu Tian. Rethinking the reliability of multi-agent system: A perspective from byzantine fault tolerance, 2025.