Surprised to kill: quantifying LLM uncertainty in morally-charged triadic dialogues

Vanessa Vanzan[†], **Nikolai Ilinykh**[†‡], **Erik Lagerstedt**[†],
Amy Han Qiu[†], **Vladislav Maraev**[†‡], **Christine Howes**[†‡]
[†]Department of Philosophy, Linguistics and Theory of Science (FLoV)
[‡]Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
name.surname@gu.se

Abstract

Multi-party dialogues on ethically and socially challenging (morally charged) topics pose a challenge for large language models (LLMs) trained on massive text corpora. Nevertheless, LLMs can illuminate features of interaction in such dialogues and serve as evaluation proxies. We propose using LLM surprisal as an indicator of points in dialogue which address or relate to the discussion of social norms on a corpus of triadic text conversations from the Balloon Task, in which three participants collaboratively resolve a moral dilemma. We hypothesise that (1) turns featuring indirect reference and implicit moral justification will exhibit higher surprisal than turns with direct reference or explicit justification, and (2) including dialogue-act or reference-type annotations in the prompt will reduce model uncertainty with the help of retrieval-augmented generation. By presenting our planned experiments, we aim to inform the design of socially aware dialogue systems able to reliably interpret nuanced ethical discourse.

1 Introduction and motivation

Large language models (LLMs) are now used across a wide range of tasks and their performance is quite good on many of them, including chat-based, game-like scenarios (Chalamalasetti et al., 2023). However, human chat can cover a variety of topics, and some discussions can be *socially charged* – they may invoke and even challenge broadly accepted social principles, for example, the norm "do not kill a child". Previous work has investigated the extent to which LLMs encode moral norms from different countries (Ramezani and Xu, 2023) and, unsurprisingly, has found that their knowledge is biased toward English-centric norms. SOCIAL-CHEM-101 (Forbes et al., 2020) provides a large-scale corpus of social norms formulated as rules of thumb, which can be used as tests of social norm understanding. More recently, Ammanabrolu et al. (2022) introduced a benchmark designed to test whether agents can act according to specified social norms during interactive scenarios, while Rao et al. (2023) showed that GPT-4 can follow explicitly prompted ethical values.

In our ongoing work we evaluate LLMs in *dialogical*, *multi-agent* settings. In these situations responses and actions are highly context-dependent, tightly interwoven, and require tracking who is in the focus of the discussion as well as the type of argument about them. We will examine how well LLMs model ethically loaded, three-participant conversations by analysing LLM surprisal on the token-/ and turn-level. We will use the Balloon Task (Lavelle et al., 2012; Howes and Lavelle, 2023), a collaborative moral dilemma in which three participants must agree on which one of four characters to sacrifice to save the others.

2 Data

Our dataset comes from the Balloon Task, a moral-dilemma discussion in which three participants interact via a text-based interface provided by the Dialogue Experimental Toolkit (DiET; Healey et al., 2003). In this dataset the server automatically inserted artificial emojis at the end of turns containing decision-related words (e.g., "kill") (Vanzan et al., 2024). Emojis were selected based on the Emoji Sentiment Ranking (Kralj Novak et al., 2015) and added every five turns, as if they had been produced by one of the participants. Importantly, they were visible only to the two recipients. Two conditions were tested: a congruent one (e.g., "kill" + an emoji typically associated with negative sentiment, such as the crying face emoji) and an incongruent one (e.g., "kill" + an emoji typically associated with positive sentiment, such as the smiling face emoji). Although the sentiment of emojis can vary with context, these associations reflect their conventional interpretations.

Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue, September 3–5, 2025, Bielefeld, Germany.

3 Proposed methodology: surprisal

Analysing linguistic data on socially charged, morally challenging topics is difficult as participants often respond implicitly rather than stating their views outright. We propose using surprisal (Hale, 2001; Levy, 2008) to flag dialogue segments that may carry heightened social or ethical weight, especially when topics and viewpoints are not expressed explicitly. Because large language models are trained on vast and diverse linguistic data, far exceeding the range and variety present in our dataset's dialogues, they encapsulate a broad spectrum of common associations (Tsimpoukelli et al., 2021). Using LLM-derived surprisal thus allows us to use the model as a proxy for identifying dialogue segments that convey non-conventional or unexpected meaning.

Formally, for a word w_t given the preceding context $w_{<t}$, surprisal is the negative log-likelihood of that word:

$$I(w_t) \;=\; -\log P_\theta\big(w_t \mid w_{<t}\big),$$

where P_θ is the probability distribution defined by the LLM. Surprisal is widely used in psycholinguistics: a word's surprisal predicts reading difficulty and correlates with processing effort (Demberg and Keller, 2008; Wilcox et al., 2023). By measuring how "surprised" an LLM is at each turn, we aim to determine whether certain discourse features such as indirect references or nuanced moral justifications systematically increase the model's uncertainty.

4 Proposed experimental design

We will segment each dialogue into individual turns, each contextualised by the preceding conversation. Surprisal will be computed at both the token and turn levels, and we will normalise it by token count to control for variation in turn length. Because particular lexical items may systematically raise or lower surprisal, we will also investigate whether high-surprisal words are linked to social-norm content or to the Balloon Task scenario itself.

Dialogue turns will be categorised based on turn-level annotations designed as follows:

- **Reference type**: explicit (e.g., "the doctor") versus implicit (e.g., "she") references to dilemma characters.
- **Argument type**: explicit moral justification (e.g., "We should eliminate the doctor because her research is useless.") versus implicit moral justification (e.g., "She could still be useful.").

In this categorisation explicit refers to expressions whose referent is clear within the current turn, without requiring prior discourse or task context. Implicit refers to expressions whose interpretation depends on such prior context.

For evaluation, we will compare the mean surprisal scores produced by the LLM across the previously defined categories, examining differences between turns annotated as explicit versus implicit. We will also analyse the temporal evolution of surprisal within each dialogue to identify patterns that may be indicative of participants' decision-making processes.

5 Final remarks

In this exploratory study, our goal is to investigate token-/ and turn-level surprisal as a proxy for LLM uncertainty in morally complex, triadic dialogues. We will also explore whether LLMs' surprisal can help identify turns involving (non-conventional) moral discussions in text-based interactions.

Although our experiments are forthcoming, we anticipate that analyses on our Balloon Task corpus will provide us with insights consistent with the following hypotheses:

- **H1**: Turns featuring implicit references and moral justifications are expected to exhibit higher surprisal because of their greater contextual complexity.
- **H2**: Including explicit contextual annotations in prompts should lower surprisal, indicating reduced model uncertainty. We aim to test this hypothesis by using retrieval-augmented generation (Lewis et al., 2020) to provide an LLM with more explicit content which is supposed to lower its uncertainty and surprisal.

Once these evaluations are complete, we will interpret how shifts in surprisal correspond to specific discourse features (including use of statistical testing) and assess the efficacy of annotation strategies. Ultimately, our goal is to inform the design of socially aware dialogue systems that can transparently and reliably engage with ethically charged content. Future directions include exploring alternative uncertainty metrics such as entropy (Shannon, 1948), testing additional annotation schemas, and integrating these techniques into interactive moral-decision support tools.

Acknowledgments

This research was supported by ERC Starting Grant DivCon: Divergence and convergence in dialogue: The dynamic management of mismatches (101077927) and by the Swedish Research Council grant (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. Vladislav Maraev was supported by Swedish Research Council (VR) grant 2023-00358 – Social laughter for virtual agents (SocLaVA).

References

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, Seattle, United States. Association for Computational Linguistics.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219, Singapore. Association for Computational Linguistics.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Patrick G. T. Healey, Matthew Purver, Jonathan King, Jonathan Ginzburg, and Gregory J. Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 539–544.

Christine Howes and Mary Lavelle. 2023. Quirky conversations: how people with a diagnosis of schizophrenia do dialogue differently. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1875):20210480.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PLOS ONE*, 10(12):e0144296.

Mary Lavelle, Patrick G. T. Healey, and Rosemarie McCabe. 2012. Is nonverbal communication disrupted in interactions involving patients with schizophrenia? In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 1772–1777.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc.

Vanessa Vanzan, Amy Han Qiu, Fahima Ayub Khan, Chara Soupiona, and Christine Howes. 2024. Emoji-text mismatches: Stirring the pot of online conversations. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Trento, Italy. SEMDIAL.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.