

794 A Limitations

795 **Noise-free Setting:** The experiments are conducted on clean, curated datasets without considering
796 real-world noise. These factors may significantly impact the robustness of detection models when
797 deployed in practical settings.

798 **Well-specified Model:** We assume that pre-trained language models used in our benchmarking
799 are well-suited for the detection task. However, these models were originally trained for language
800 generation rather than detection, and suboptimal fine-tuning or domain mismatch may limit their
801 effectiveness in distinguishing H-H and H-C dialogue.

802 **Asymptotic Approximations:** Some of the statistical analysis techniques employed rely on asymp-
803 totic assumptions that require large sample sizes to achieve accurate estimation. In practice, especially
804 with limited or imbalanced datasets, these approximations may not hold, potentially affecting the
805 validity of the results.

806 B Implementation Details

807 B.1 Hardware devices

808 All our experiments were meticulously conducted on a high-performance computing platform running
809 Ubuntu. The platform is powered by an Intel(R) Xeon(R) Platinum 8176 CPU @ 2.10GHz, delivering
810 robust computational capabilities. The system is equipped with a substantial 503 GB of memory,
811 ensuring efficient data processing and storage. Additionally, to further enhance computational power,
812 we utilized four NVIDIA Corporation GA102GL RTX A6000 GPUs. These GPUs provided the
813 necessary parallel processing power to handle the intensive computational tasks associated with our
814 research. The stability and broad support of the Ubuntu operating system allowed us to fully leverage
815 the hardware’s performance, ensuring the smooth execution of experiments and the reliability of our
816 results.

817 B.2 Datasets

- 818 • **DailyDialog:** This dataset contains high-quality, multi-turn dialogues reflecting everyday
819 human communication. The conversations cover various topics and exhibit natural lan-
820 guage usage. We utilized dialogues directly from this corpus as part of our H-H chit-chat
821 data, selecting conversations exceeding a minimum turn length threshold suitable for GCT
822 analysis.
- 823 • **PersonaChat:** This dataset consists of chit-chat dialogues where participants are assigned
824 specific persona profiles that they are expected to condition their conversation on. It
825 encourages engaging and consistent dialogue. Similar to DailyDialog, naturally occurring
826 dialogues between human participants in this dataset were included in our H-H chit-chat
827 corpus.
- 828 • **MultiWOZ :** A large-scale, multi-domain dataset for task-oriented dialogues, covering
829 domains like restaurants, hotels, transportation, etc. It is a standard benchmark in dialogue
830 state tracking and end-to-end dialogue systems. We used the human-human Wizard-of-
831 Oz collected dialogues within this dataset, where one human plays the user and another
832 simulates a constrained system based on database information, as representative examples
833 for our H-H task-oriented corpus.
- 834 • **Taskmaster-1:** This dataset contains goal-oriented dialogues, covering tasks such as ordering
835 pizza, creating auto repair appointments, and booking flights. It includes both spoken and
836 written conversations collected via a Wizard-of-Oz setup. We specifically used the written
837 dialogues where both the ‘user’ and the ‘wizard’ (simulating the system) were human
838 participants to form part of our H-H task-oriented corpus.

839 **H-H Corpus:** The Human-Human (H-H) corpus used in our experiments was formed by selecting
840 dialogues directly from the aforementioned datasets (DailyDialog, PersonaChat, the human-controlled
841 segments of MultiWOZ, and Taskmaster-1 WoZ data). Dialogues below a pre-defined length (e.g., 20
842 turns) were filtered out to ensure suitability for Granger Causality analysis.

H-M Corpus Construction: To create a comparable Human-LLM (H-M) corpus, we recruited human participants (e.g., via crowd-sourcing or internal volunteers) and tasked them with interacting with a specific state-of-the-art LLM (e.g., specify the model used, like Llama-2-Chat 70B or GPT-4 via API, ensuring compliance with terms of service). Crucially, the interactions were designed to mirror the contexts of the H-H datasets:

For chit-chat, participants were encouraged to discuss topics similar to those found in DailyDialog or were assigned personas based on the PersonaChat dataset to interact with the LLM. For task-oriented dialogues, participants were given specific goals identical or analogous to those in the MultiWOZ and Taskmaster-1 datasets (e.g., "Find and book a moderately priced Italian restaurant for two people") and instructed to achieve these goals solely through interaction with the designated LLM.

This process ensured that the H-M dialogues, while involving an AI agent, covered similar domains, intents, and task structures as the H-H dialogues, allowing for a more controlled comparison of interaction dynamics.

H-M Corpus Construction: The Human-LLM (H-M) corpus was constructed semi-synthetically, derived directly from the dialogues selected for the H-H corpus to ensure maximal comparability of user input and conversational context. For each selected H-H dialogue $C^{(i)} = \{(U_1, A_1), (U_2, A_2), \dots\}$ (where U_t denotes a user utterance and A_t denotes the original human agent’s utterance at turn t), we identified all turns originally spoken by the human agent A . We then employed a specific pre-trained Large Language Model (LLM-X, e.g., specify model like Llama-2-Chat 70B or GPT-4) to generate alternative responses for these turns.

Specifically, for each agent turn A_t , the dialogue history preceding it, typically ending with the user’s utterance U_t , was provided as context to LLM-X. Let the history be $H_t = (U_1, A'_1, U_2, A'_2, \dots, A'_{t-1}, U_t)$ where A'_k are the previously generated LLM responses (or original A_k for $k = 1$ if the agent starts). The LLM was prompted to generate a suitable response A'_t given this history:

$$A'_t = \text{LLM-X}(H_t) \quad (13)$$

This generated response A'_t then replaced the original human response A_t in the dialogue sequence. This process was repeated for all agent turns in the dialogue, resulting in a new H-M dialogue $C'^{(i)} = \{(U_1, A'_1), (U_2, A'_2), \dots\}$. Note that the user utterances U_t remain identical to those in the original H-H dialogue $C^{(i)}$.

For dialogues derived from PersonaChat, the corresponding persona information was included in the prompt for LLM-X to encourage consistent persona adoption. For task-oriented dialogues derived from MultiWOZ and Taskmaster-1, relevant task goals or simulated dialogue states (if available and applicable) were potentially included in the prompt history H_t to guide the LLM towards task completion, mimicking the information available to the original human agent/wizard. This construction method yields an H-M corpus where the user’s side of the conversation is natural human language drawn from established datasets, while the agent’s side is generated by the target LLM conditioned on that human input, allowing for a controlled comparison of response patterns and interaction dynamics against the original H-H dialogues. Similar length filtering was applied to the resulting H-M dialogues.

B.3 Metrics

To ensure the accuracy and reliability of the results, each experiment was conducted in triplicate, and the standard deviations were calculated. This approach effectively assesses the stability and consistency of the data, thereby enhancing the credibility of our conclusions. To assess the detector’s capability to differentiate between texts generated by large language models (LLMs) and those written by humans, we utilize Accuracy (A) and the Area Under the Receiver Operating Characteristic Curve (AUROC) as primary performance metrics. Additionally, we consider other metrics, such as F1 scores (F1) and Recall (R), to provide a more comprehensive evaluation.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

$$R = \frac{TP}{TP + FN}; \quad F1 = 2 \times \frac{P \times R}{P + R} \quad (15)$$

891 True Positives (*TP*) refer to H-H dialogue correctly identified by the model. True Negatives (*TN*)
892 represent H-C dialogue accurately classified as H-c dialogue. False Positives (*FP*) denote H-C
893 dialogue incorrectly labeled H-H dialogue, while False Negatives (*FN*) correspond to H-H dialogue
894 the model fails to identify correctly.

895 **C Potential Positive Societal Impacts**

896 **Enhanced Dialogue Understanding and Interaction:** By leveraging interaction dynamics and
897 semantic-focused attribution, this research aims to improve dialogue understanding and classification
898 accuracy beyond purely semantic analysis. This could lead to more effective communication tools,
899 such as chatbots and virtual assistants, enhancing user experience and satisfaction across various
900 applications.

901 **Improved Detection of AI-Generated Text:** The development of sophisticated models for detecting
902 machine-generated text can play a crucial role in combating misinformation and ensuring content
903 authenticity. In an era where AI-generated content is becoming increasingly prevalent, having reliable
904 methods to distinguish between human and AI-generated texts is vital for maintaining trust in digital
905 communications.

906 **Promotion of Ethical Use of AI:** Through advancements in identifying AI-generated content, this
907 research supports the ethical use of technology by helping prevent misuse and manipulation. It
908 contributes to the broader conversation on AI ethics and responsibility, encouraging transparency and
909 accountability in how AI technologies are deployed and managed.