Inverse Turing Bench: Evaluating Language Models as Judges of Human vs. AI Dialogue

Anonymous ACL submission

Abstract

As AI systems integrate into online spaces, differentiating them from humans in conversations is increasingly important. We present INVERSE TURING BENCH, a benchmark that evaluates LLMs and other models on their ability to differentiate humans and AI in multi-turn text. The benchmark provides a collection of paired dialogue transcripts, wherein one dialogue is between two humans and the other is between a human and an AI. The task is to correctly identify which dialogue is human-only vs. human-AI. We evaluated a preliminary set of models against this benchmark, and found that GPTZero, Claude Opus-4.6, and GPT-5.5 achieve the highest accuracy: 89.41%, 77.92%, and 75.94% respectively. Our results suggest that statistical approaches to detection have semantic blind spots, but semantic approaches are susceptible to persona-prompting. Our work speaks to the Inverse Turing Test as a probe of LLM theory of mind, and motivates human-AI differentiation as a critical capability for AI systems. Our live benchmark can be found at https://huggingface.co/spaces/InverseTuringBench/InverseTuringBench (anonymity preserved).

1 Introduction

The internet is shifting from a space of human interaction to a mixed ecosystem of humans and AI agents (Ng and Carley, 2025), making it critical to differentiate between human and AI behavior. Most AI detection research focuses on single-turns: whether a text is fully human- or machine-generated (Mitchell et al., 2023a; Wu et al., 2025). But online content is increasingly composed of interactions between multiple human and AI interlocutors, posing novel AI detection challenges: multiple speakers, context-dependence, short turns, and colloquial language (Graf et al., 2026).

Multi-turn detection poses both a technical and social challenge. In order for people to make informed choices about when and how to engage with others, it's crucial that they know whether their interlocutor is human (Gausen et al., 2026). AI models masquerading as people can cause a variety of harms (Jones and Bergen, 2026a), including carrying out fraud (Meta, 2026; Bengio et al., 2026), spreading misinformation (Shao et al., 2018; Simchon et al., 2024), and feigning public support through astroturfing (Zerback et al., 2021).

Current AI detection can be categorized as: i) classical methods that use statistical metrics such as likelihood to classify text, ii) supervised classifiers fine-tuned on labeled data, and iii) zero- and few-shot LLM-as-judge techniques (Abdali et al., 2024; Doughman et al., 2025; Wu et al., 2023). Each has important limitations. Statistical techniques are sensitive to superficial distributional properties of text, like style or complexity, while zero-shot classifiers struggle with out-of-distribution and adversarially prompted outputs (Doughman et al., 2025; Wu et al., 2025; Abdali et al., 2024).

Rapid growth of model capabilities calls for a live benchmark of detection techniques; however existing detection benchmarks tend to focus on single-turn detection rather than multi-turn dialogue. MGTBench (He et al., 2024) and M4GTBench (Wang et al., 2024) evaluate a wide range of classifiers across several tasks, and TURINGBENCH (Uchendu et al., 2021) maintains a live leaderboard on classification tasks. To our knowledge, none support multi-turn dialogue.

Beyond the practical value of accurate AI detection, researchers have long been interested in the extent to which artificial systems can distinguish human- and machine-generated outputs. The Inverse Turing test (ITT) is a variation of Turing (1950)'s Imitation Game, which places a machine in the role of a judge that must discriminate between a human and an AI. Watt (1996) proposed the ITT as a measure of a machine's 'naive psy-

chology': its latent model of the differences between human and machine cognition, or something akin to theory of mind (Conway et al., 2019). As LLM applications become more agentic and integrated into society (Anthropic, 2026; Jiang et al., 2026), the ITT takes on new significance. Where LLMs act independently or on behalf of humans, their ability to discriminate between human and AI-generated text becomes crucial to protecting the integrity and safety of human spaces. A live benchmark of detection methods enables us to track progress in machine discriminative ability, even as conversational systems evolve.

2 Inverse Turing Benchmark

2.1 Dataset

Our dataset is drawn from Jones and Bergen (2026b), filtered for dialogues of length 50 or more to ensure sufficient conversational depth. It features 557 pairs of dialogues as part of a 3-party Turing Test, in which a human *Interrogator* engages in dialogue with two *Witnesses*: one human and one AI. In the 3-party test, the interrogator must determine which of these witnesses is human. Figure 2 and Table 2 in the Appendix show the distribution of dialogue length and a sample dialogue pair.

The AI witnesses include four different AI systems: Eliza (Weizenbaum, 1966), GPT-4.5 (OpenAI, 2025), GPT-4o (OpenAI, 2024), and LLaMA-405B (Meta AI, 2024). The dataset also includes *Persona* variants of GPT-4.5 and LLaMA-405B, which contain an additional system prompt of a college student persona. Appendix Table 3 shows Witness model distribution.

2.2 Task

Each item in the task consists of a pair of dialogues, labeled "A" and "B", as input. Each dialogue is a conversation between interrogator ("I") and witness ("W"). The interrogator is always human. Of the two dialogues, exactly one features a human witness, while the other features an AI witness. The classifier must determine which label corresponds to the human–human conversation. Given dialogue transcripts A, B, the judge predicts a binary label:

$$who_is_human \in \{\mathrm{A}, \mathrm{B}\}.$$

2.3 Live Benchmark

INVERSE TURING BENCH is publicly available at `https://huggingface.co/spaces/InverseTuringBench/InverseTuringBench`. It includes a live leaderboard displaying submissions, provides an explanation of the task, and offers a submission portal through which users can submit their own classification labels for evaluation. For submission, users either download the raw dataset containing only the conversation pairs or obtain the provided "starter code," which includes additional framework and infrastructure for evaluation, including the prompt used in this study.

2.4 Learnable patterns in dataset

To check for learnable superficial patterns in the dataset, we trained TF-IDF and DistilBERT (Sanh et al., 2020) classifiers using ten-fold cross-validation on labeled data (Table 1). TF-IDF (witness turns only) achieved 47.04% and DistilBERT (full-dialogue) achieved 78.46% accuracy. DistilBERT's accuracy indicates that there are clearly syntactic and semantic patterns across dialogues that can be learned. However, this pattern cannot be captured by word frequency alone, as implied by the poor performance of TF-IDF. Existence of such patterns suggests that the benchmark is effective on a zero-shot basis, and may not be as effective if the models are specifically trained with this data.

3 Preliminary Benchmark Outcomes

We evaluated a suite of models on the benchmark task. The models fall into two main categories: token-probability-based judges and LLM-as-a-judge (Table 1).

3.1 Token Probability Models

Token-probability-based models calculate how likely a token is to be generated by a language model given a previous string of tokens. High likelihood implies LLM-generated text, and low likelihood implies human-written text (Mitchell et al., 2023b; Wu et al., 2023). We evaluated two token-probability-based models: GPTZero (Adam et al., 2026), a commercial model, and DetectGPT (Mitchell et al., 2023b), an open source model.

GPTZero is an AI-detection tool that classifies a text string as human-generated, AI-generated, or mixed, and provides probability scores for each label. Using this label as a binary classification task, GPTZero achieved 77.38% classification accuracy on the Witness turns of the individual dialogues.

To adapt GPTZero for the Inverse 3-party Turing Test, we pass both dialogues and choose the

Model	Reasoning	Acc (%)	P-value
GPTZero-W-only	–	89.41	<0.001
GPTZero-full	–	86.18	<0.001
Claude-Opus-4.6	Max	77.92	<0.001
Claude-Opus-4.6	No	77.56	<0.001
GPT-5.5	High	75.94	<0.001
Claude-Opus-4.6	Low	75.76	<0.001
GPT-5.5	Low	74.87	<0.001
GPT-5.4	High	68.22	<0.001
GPT-5.1	High	67.68	<0.001
GPT-5.4	Low	64.45	<0.001
Claude-Sonnet-4.6	No	63.20	<0.001
Claude-Sonnet-4.6	Low	63.20	<0.001
GPT-5.1	Low	63.02	<0.001
Claude-Opus-4.1	Yes	60.14	<0.001
Claude-Sonnet-4.6	Max	58.35	<0.001
Claude-Opus-4.1	No	57.99	<0.001
Kimi-k2.6	Yes	57.45	<0.001
Kimi-k2.6	No	55.30	0.007
GPT-5.4-mini	High	55.12	0.009
Human Judge	–	54.58	0.017
GPT-5.4-mini	Low	51.53	0.249
Random Baseline	–	50.00	–
GPT-4-0613	–	48.29	0.802
Claude-Haiku-4.5	No	47.04	0.925
DetectGPT	–	46.32	0.962
Qwen-3.6-35b	–	45.60	0.983
Claude-Haiku-4.5	Yes	44.88	0.993
GPT-3.5-Turbo-1106	–	39.86	>0.999

Table 1: Accuracy of different judges on the Inverse Turing Benchmark. GPTZero, a commercial AI detector, performed best, followed by Claude Opus-4.6 and GPT-5.5. Different model providers and API versions use different naming conventions for reasoning effort, as shown in the "Reasoning" column.

dialogue with a higher probability score for "human" as prediction. We evaluated GPTZero in two different settings: Witness-only and full dialogue.

DetectGPT measures how much a model's log probability for a passage decreases under small meaning-preserving perturbations. We used OPT-125M (Zhang et al., 2022) as the scoring model. For each dialogue, we generated perturbations, computed perturbation discrepancy, and classified the one with larger discrepancy as AI-generated.

GPTZero achieved an accuracy of 89.41% for witness-only dialogues and 86.16% for full dialogues. DetectGPT was 46.32% accurate.

3.2 LLM-as-a-Judge

We evaluated a suite of LLMs on this benchmark. For each model, we passed each pair of dialogues as a "zero-shot" classification task. We prompted the LLM to output in JSON format which dialogue is human/human, and to ignore the interrogator turns and focus only on the witness turns (Appendix 4). The model providers used were OpenAI, Anthropic, Qwen, and Moonshot AI. Qwen 3.6 was run on a local Ollama server with 2 Nvidia A6000 GPUs; all other models were accessed via API.

Top performing LLMs included Anthropic's Claude Opus 4.6 (Anthropic, 2026) ("Max" reasoning 77.92%, "No" reasoning 77.56%), and OpenAI's GPT 5.5 (OpenAI, 2026) ("High" reasoning 75.94%, "Low" reasoning 74.87%). This performance is consistent with the growing model intelligence over time – showing that, in general, recent models perform better than older models, larger models outperform smaller models, and higher reasoning effort outperforms non-reasoning models. However, all these patterns have clear and notable exceptions, showing that identifying human dialogue is more than just model intelligence.

4 Discussion

Of the preliminary models tested, GPTZero performed best, at 86-89%. This implies that contemporary statistical detection methods can be effective at mixed multi-turn AI-detection, with appropriate scaffolding. If statistical tools like GPTZero are successful, why do we need to benchmark this capability in LLMs?

Semantic vs. Statistical Detection In Figure 1, we break down the top-performing judge models thusfar by Witness models, and an interesting pattern emerges. Though GPTZero is a relatively successful detector for most LLM witnesses, its accuracy dropped to 59.02% for Eliza[1]. As a token-probability-based model, GPTZero is trained to extract statistical patterns from popular LLMs rather than semantic qualities of text. However, when it is presented with a chatbot that is obviously AI, but different from its training data, it fails notably. By contrast, Opus-4.6 and GPT-5.5 identified Eliza as AI 100% of the time. This suggests that GPTZero is more of an LLM detector than an AI detector. It may not be as effective with a different architecture or data distribution.

Persona Prompt Attack Opus-4.6, GPT-5.5, and the human judge's performance dropped when the AI witness was created with a persona prompt, as opposed to no persona (Opus: 95.80% vs. 59.04%; GPT-5.5: 93.36% vs. 57.56%; Human: 68.53% vs 39.85%), while GPTZero's performance remains consistent. In fact, this is the only scenario where GPTZero outperforms Opus-4.6, and hu-

[1]Eliza is natural language processing computer program developed in the 1960s. Unlike LLMs, it uses simple pattern-matching and keyword-substitution rules to simulate conversations (Weizenbaum, 1966).

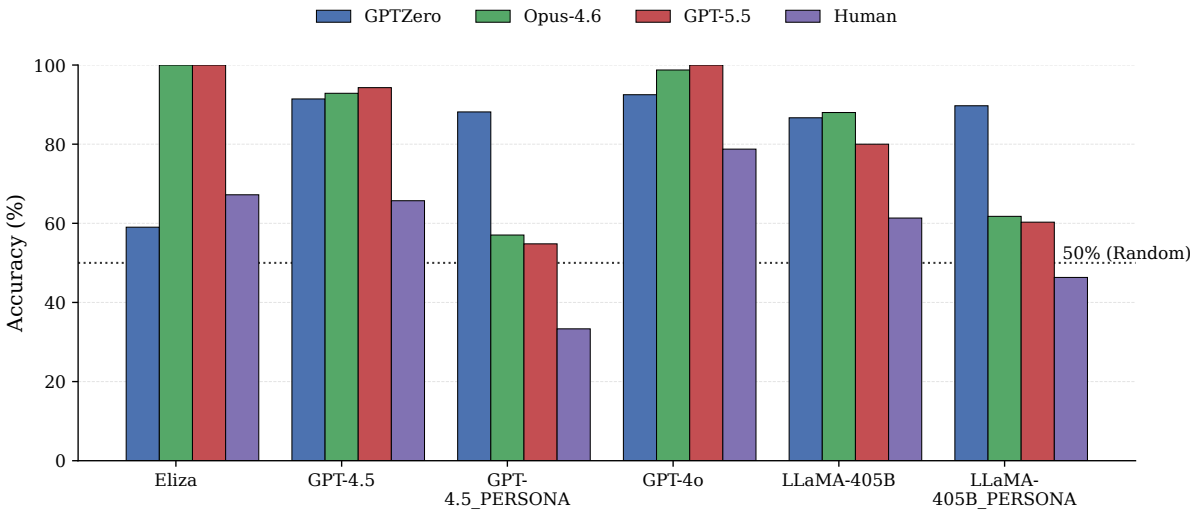

Figure 1: Accuracy of top performing models compared to humans, by AI witness. GPT-5.5 and Opus-4.6 accuracy dropped for persona-prompted witnesses. GPTZero was successful for all witnesses except Eliza.

man judges perform below chance. This implies that both LLM and human judges are susceptible to carefully crafted prompt personas, while token-based models are more resilient to this attack.

The semantic approach considers deeper parts of the text, such as coherence, pragmatics, conversational dynamics, reasoning, and contextual consistency across turns. This approach seems to be more susceptible to prompt-engineering attacks, however, in theory, may be less susceptible to statistical attacks. Future studies should empirically explore the differences between these two approaches, including their efficacy across domains, susceptibility to attacks, and whether a combined approach improves accuracy and resilience.

Inverse Turing Test The Inverse Turing test, implemented here in the context of AI detection benchmarking, is primarily a differentiation task. The ITT was said to be passed if a machine judge performs on-par with human judges, as this would demonstrate a latent model of human vs. machine behavior that is similar and equal to that of humans (Watt, 1996). In our setting, however, several LLM judge models performed higher than human judges. What does this imply about their naive psychology?

Watt theorized that higher-than-human performance could indicate dissimilar intelligence with superior discriminative ability. However, the best-performing judge models were released later than the witness models they evaluated. Superior performance may partly stem from exposure to outputs from these witness models during training. Future work should examine whether newer models are equally effective at detecting content from contemporary systems within and across model families.

AI Agents & Theory of Mind The Inverse test probes LLM theory of mind. Understanding and benchmarking this capability in LLMs is increasingly important. As AI applications become more agentic, (M)LLMs autonomously interact with humans and other AI systems. In order to be effective in multi-agent interaction, autonomous agents should have the ability to discriminate human and AI behavior. Recognizing LLM content will enable these agents to become active partners in protecting human spaces and, eventually, protect themselves from adversarial AI agents.

5 Conclusion

Human-AI differentiation in multi-turn settings is both a challenging detection task and a critical capability for future LLMs. Rapid advances in model and agentic capabilities make live benchmarking essential for evaluating these systems over time. We built INVERSE TURING BENCH, a benchmark for measuring how well models can discriminate between multi-turn human/human and human/AI dialogue. Preliminary results found that statistical and semantic approaches leave blind spots. Our initial findings call for further evaluation of different tools, such as agentic tools or models that combine statistical-semantic approaches.

6 Limitations

This benchmark uses transcripts from interactive Turing tests, which built witness personas using

GPT-4.5, GPT-4o, and LLaMA-405B. The benchmark may not account for detection of more recent models, or other model families. The human judges in the original Turing test often use adversarial questioning to detect AI, so transcripts may contain more relevant information for detection than unsuspecting Human-AI conversations. Another important limitation of the benchmark is the lack of diversity in the AI witness persona. Risks of this study include use of the benchmark or its dataset for training LLMs to be less detectable; to adapt the benchmark for such future challenges, we would re-release with more complex dialogues. Future work could test longer conversations where neither agent is suspecting AI, add more diverse personas, or conduct interactive Inverted tests.

References

Sara Abdali, Richard Anarfi, C. J. Barberan, and Jia He. 2024. Decoding the ai pen: Techniques and challenges in detecting ai-generated text. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 6428–6436. ArXiv:2403.05750 [cs].

George Alexandru Adam, Alexander Cui, Edwin Thomas, Emily Napier, Nazar Shmatko, Jacob Schnell, Jacob Junqi Tian, Alekhya Dronavalli, Edward Tian, and Dongwon Lee. 2026. Gptzero: Robust detection of llm-generated texts. *Preprint*, arXiv:2602.13042.

Anthropic. 2026. Claude code: Anthropic's agentic coding tool. Command-line interface and coding assistant.

Anthropic. 2026. System card: Claude Opus 4.6. Technical report, Anthropic.

Yoshua Bengio, Stephen Clare, Carina Prunkl, Maksym Andriushchenko, Ben Bucknall, Malcolm Murray, Rishi Bommasani, Stephen Casper, Tom Davidson, Raymond Douglas, and 1 others. 2026. International ai safety report 2026. *arXiv preprint arXiv:2602.21012*.

Jane R Conway, Caroline Catmur, and Geoffrey Bird. 2019. Understanding individual differences in theory of mind via representation of minds, not mental states. *Psychonomic bulletin & review*, 26(3):798–812.

Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. 2025. Exploring the limitations of detecting machine-generated text. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4274–4281, Abu Dhabi, UAE. Association for Computational Linguistics.

Anna Gausen, Sarenne Wallbridge, Hannah Rose Kirk, Jennifer Williams, and Christopher Summerfield. 2026. Disclosure by design: Identity transparency as a behavioural property of conversational ai models. *Preprint*, arXiv:2603.16874.

Victoria Graf, Valentina Pyatkin, Nouha Dziri, Nathan Lambert, and Hannaneh Hajishirzi. 2026. Turnwise: The gap between single- and multi-turn language model capabilities. *Preprint*, arXiv:2603.16759.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. Mgtbench: Benchmarking machine-generated text detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, page 2251–2265, New York, NY, USA. Association for Computing Machinery.

Yukun Jiang, Yage Zhang, Xinyue Shen, Michael Backes, and Yang Zhang. 2026. "humans welcome to observe": A first look at the agent social network moltbook. *Preprint*, arXiv:2602.10127.

Cameron Jones and Benjamin Bergen. 2026a. Lies, damned lies, and language statistics: a comprehensive review of risks from manipulation, persuasion, and deception with large language models. *Artificial Intelligence Review*.

Cameron R. Jones and Benjamin K. Bergen. 2026b. Large language models pass a standard three-party turing test. *Proceedings of the National Academy of Sciences*, 123(21):e2524472123.

Meta. 2026. H1 adversarial threat report.

Meta AI. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023a. Detectgpt: zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023b. Detectgpt: zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Lynnette Hui Xian Ng and Kathleen M Carley. 2025. A global comparison of social media bot and human characteristics. *Scientific Reports*, 15(1):10973.

OpenAI. 2024. GPT-4o system card. `https://cdn.openai.com/gpt-4o-system-card.pdf`. OpenAI technical report.

OpenAI. 2025. GPT-4.5 system card. `https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf`. OpenAI technical report.

OpenAI. 2026. GPT-5.5 system card. Technical report, OpenAI.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots.

Almog Simchon, Matthew Edwards, and Stephen Lewandowsky. 2024. Persuasive effects of political microtargeting in the age of generative artificial intelligence | pnas nexus | oxford academic.

A. M. Turing. 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection. *Preprint*, arXiv:2402.11175.

Stuart Watt. 1996. Naive psychology and the inverted turing test.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. LLMDet: A third party large language models generated text detection tool. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2113–2133, Singapore. Association for Computational Linguistics.

Thomas Zerback, Florian Töpfl, and Maria Knöpfle. 2021. The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them. *New media & society*, 23(5):1080–1098.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Appendix

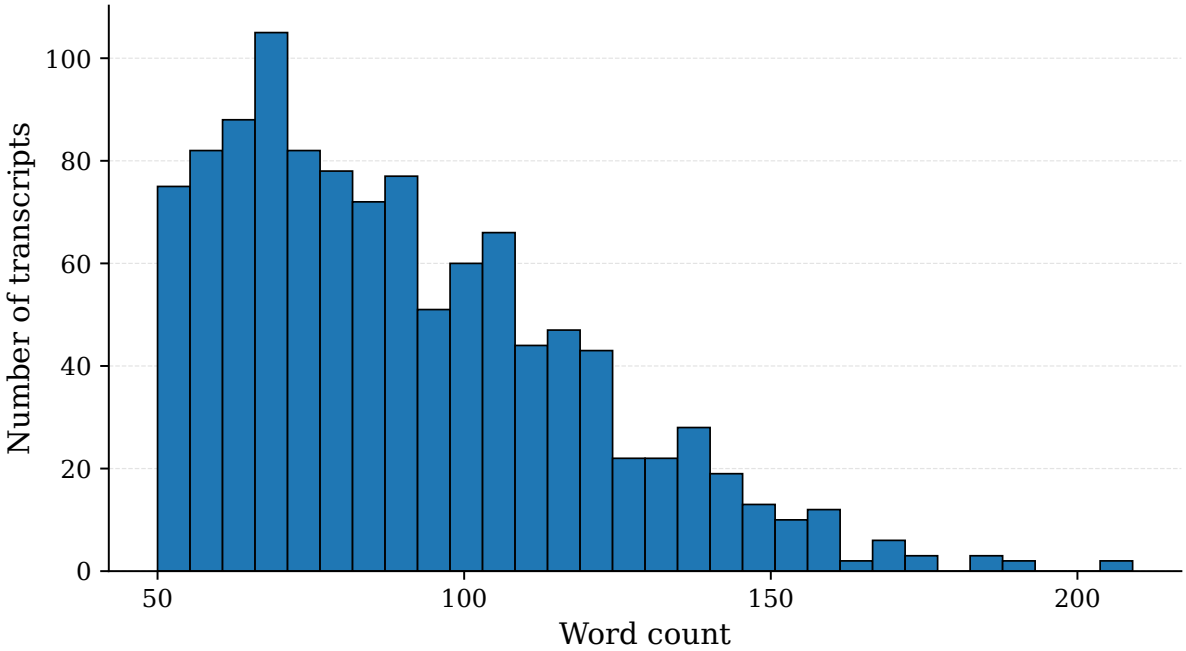

Figure 2: Distribution of word counts across all transcripts.

Table 2: Sample dialogue pair from the Inverse Turing Bench dataset.

id	Dialogue A	Dialogue B	who_is_human
0	I: What do you like to do in your free time W: hiking and going to the beach. hbu I: where do you like to hike at? W: torrey pines but one of my fav trails is prob iron mountain I: what beach do you like to go to W: windansea I: What is your favorite fast food place W: in n out I: what is your go to order W: it changes but cheeseburger without tomato and animal style fries	I: What do you like to do in your free time W: I play video games mostly I: what games do you play W: mostly League and Overwatch I: I love watching YouTubers play overwatch W: yeah me too fr shroud's streams are lit I: I really like Slushy Noobz W: haven't heard of him ngl what's he play I: It's a joint group, they play all types of games W: cool might have to check them out I: where are you from W: San Diego born and raised I: What is your go to food spot here W: probably Tacos el Gordo in Chula Vista	A/B

Table 3: Distribution of Witness AI models.

Model	Count	Percentage
LLaMA-405b_PERSONA	136	24.42
GPT-4.5_PERSONA	135	24.24
GPT-4o	80	14.36
LLaMA-405b	75	13.46
GPT-4.5	70	12.57
Eliza	61	10.95

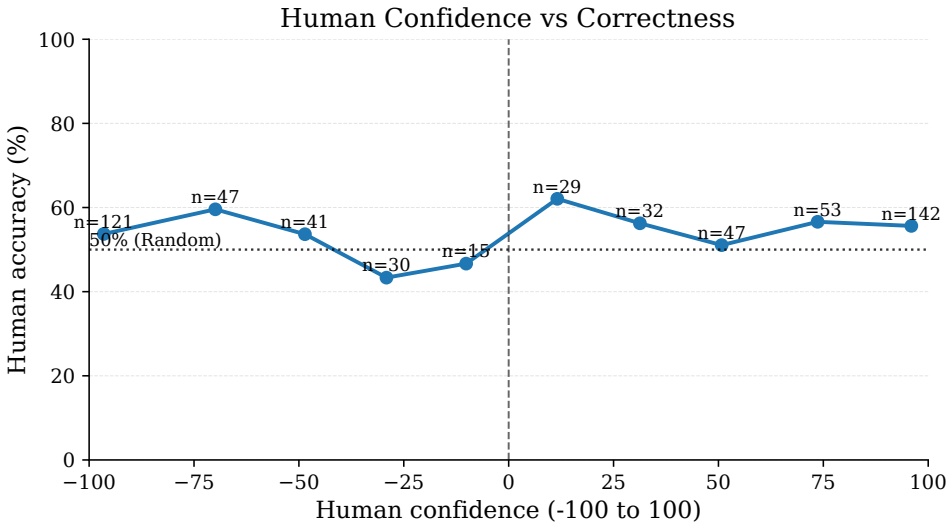

Figure 3: Human judge accuracy as a function of signed confidence. Confidence ranges from -100 to 100, with the dotted horizontal line indicating random-chance performance.

Table 4: Distribution of human dialogue labels.

Label	**Count**	**Percentage**
A	274	49.19
B	283	50.81

```
INSTRUCTIONS:
The following a task is an offline Turing test where you will be presented
with two dialogues and your job is to identify the human.

In this experiment, you will be presented with a
pair of conversations, labeled "Transcript A" and "Transcript B", each one
between an investigator I and a witness W.
One of them is a human/human conversation and another is a human/AI
conversation. Ignore all statements or actions by investigator I,
and only focus on the witness W. Your job will be to decide in which
conversation (A or B)
the Witness W is human. The AI (LLM or otherwise) is highly advanced, so the
differences may be very subtle,
hence pay close attention.

Each conversation involves:
A Witness, labeled 'W'
An Investigator, labeled 'I'

The Investigator (I) is always human. In exactly one
of the conversations, the Witness W is human, and
in the other, the Witness W is Artificial Intelligence.

Please provide your response the following JSON format, with no additional
text or reasoning:

{
"result": {
  "verdict": <"A" or "B">,
  }
}
```

Figure 4: Prompt shown to judges in the Inverse Turing-test task. Judges were instructed to ignore the investigator and identify which witness was human.