

📍MT-PINGEVAL: Evaluating Multi-Turn Collaboration with Private Information Games

Jacob Eisenstein Fantine Huot Adam Fisch Jonathan Berant Mirella Lapata

Google DeepMind

{jeisenstein,fantinehuot,fisch,joberant,lapata}@google.com

Abstract

We present a scalable and verifiable methodology for evaluating language models in multi-turn interactions, using a suite of collaborative games that require effective communication about private information. This enables an interactive scaling analysis, in which a fixed token budget is divided over a variable number of turns. We find that language models often fail to use interactive collaboration to improve over the non-interactive baseline in which one agent summarizes its information and the other agent immediately acts, despite substantial headroom. This suggests that state-of-the-art models still suffer from significant weaknesses in planning and executing multi-turn collaborative conversations. We analyze the linguistic features of these dialogues, assessing the roles of sycophancy, information density, and discourse coherence. While there is no single linguistic explanation for the collaborative weaknesses of contemporary language models, we note that humans achieve comparable task success at superior token efficiency by producing more coherent dialogues. The proactive management of private information is a defining feature of real-world communication, and this work is designed to drive further progress on this capability.

1 Introduction

Multi-turn conversation is a core capability for language models, but it is difficult to measure. The primary challenge lies in creating realistic conditions in which (1) dialogue is necessary, yet (2) performance is verifiable. Many existing evaluations focus on scripted scenarios where a human participant (or more often, a user simulator that is backed by another language model) is given a predefined goal or set of goals, and is asked to play out a multi-turn interaction with an AI assistant (e.g., Shao et al., 2024; Laban et al., 2025). There are at least two major drawbacks to this style of evaluation. First, it is fundamentally asymmetric: the assistant generates a series of proposals (e.g., a travel itinerary, some software code) that the human critiques until reaching a satisfactory output. In contrast, human-human interactions usually require *both* participants to proactively shape the direction of the dialogue. This symmetry gives rise to two of the central tasks in communication: determining what parts of one’s knowledge to share and what additional knowledge needs to be elicited from the other participant(s) (Grice, 1975; Lin et al., 2024). Second, this style of evaluation requires a faithful simulator of human users. This is a challenging task in its own right because real human speakers may enter a conversation with only a vague sense of their own goals, preferences, and constraints (Lin & Tomlin, 2025).

To measure the ability of language models to address the fundamental communicative challenges of multi-turn dialogue, we present a new form of multi-turn evaluation through collaborative *private information games* (PINGs). The key feature is that each participant has access to private information, some of which is essential to the successful completion of the task. The participants must then collaboratively solve the task by building a shared understanding of their private information through linguistic communication. To prevent participants from simply giving away their private information, we use information modalities that are not easy to convert into text (under reasonable length constraints): images and

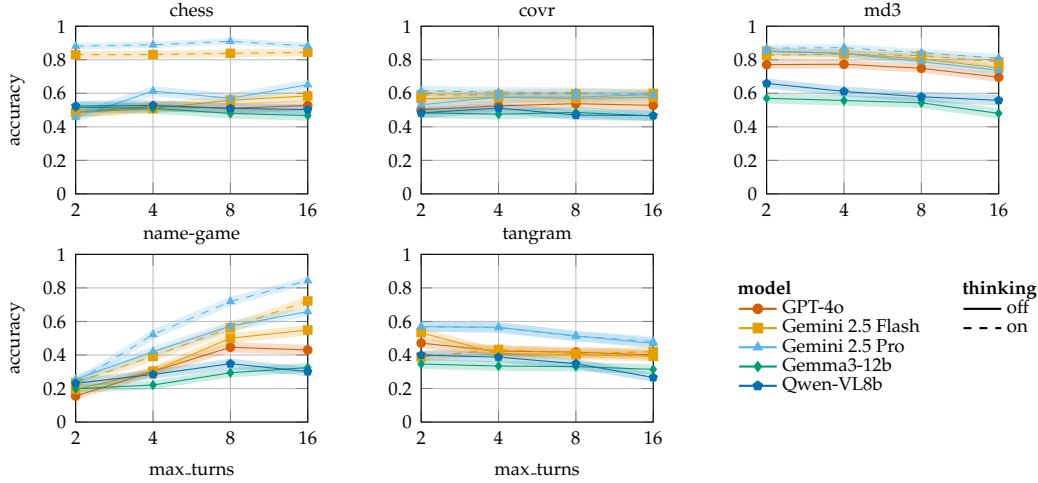


Figure 1: **Isotoken evaluation**, showing task accuracy with a constant token budget sharded over varying numbers of turns. In most cases, performance is flat or decreasing as the turn budget increases, indicating that the models fail to exploit interactivity.

structured knowledge. This basic idea echoes much earlier work on “information gap” tasks in second-language instruction, in which language learners must share private information to jointly accomplish a goal (Doughty & Pica, 1986). We collect these private information games into a multi-turn benchmark called MT-PINGEVAL. To enable comparative analysis at scale, we design the games to enable automatic verification and generation of new instances. This makes it possible to evaluate the multi-turn capabilities of language models from a new perspective that focuses on their proactive conversational capabilities.

Success or failure in multi-turn private information games depends on two capabilities: (1) interacting effectively across multi-turn dialogues; and (2) reasoning about the game itself. Our primary interest is in the interactive capability; to isolate it, we propose a novel form of multi-turn scaling analysis, in which we fix the total number of tokens that can be exchanged in the interaction, and vary the number of turns (see Section 4). The performance with two turns is then a baseline, because if given a larger number of turns, the players can achieve the same communication by ignoring each other and sharding their message from the two-turn variant into however many turns are permitted. In well-designed tasks (see formal characterization in Section 2), the players can do better by refining their communication based on the dialogue context. However, we observe that contemporary language models generally fail to achieve this. In most cases, there is no improvement as the turn budget grows, showing that the language model players are unable to use interaction to improve their task performance. In several cases, performance significantly *decreases* with higher turn budgets, because the players prematurely terminate the dialogue rather than verifying their conclusions (Cemri et al., 2025, show a related finding in practical deployment scenarios). These negative results hold broadly across model scales and thinking budgets, suggesting that there is still significant room for improvement in multi-turn communication.

2 Levels of interactivity

To formalize intuitions about the role of private information in multi-turn communication, we propose a framework for quantifying the *level of interactivity* for a collaborative private information game. Consider a two-player game, where player 1 receives private information X_1 and player 2 receives private information X_2 , which is at least partially disjoint. The shared goal is for player 1 to give an answer that gets a high score. (For this theoretical formulation it is convenient to allow only player 1 to give an answer.) Assume each player has access to an encoding function $f_i \in \mathcal{F}_i^n$, which maps the dialogue history and their

Alice	Bob	
. . b . . k r .	r n b q k b r .	1. Alice: Hi Bob. My board has the White King on g1 and a Rook on b1. The Black King is on f8. Both sides have 7
r . . q p p . .	. . p p p p p .	2. Bob: Okay Alice. My board has 8 white pawns and 8 black pawns. White King is on e1, Black King on e8.
n p p	p p p	3. Alice: Okay Bob. My board has only 6 white pawns and 7 black pawns. My White King has castled kingside to g1.
p B p		4. Bob: Okay Alice. My Kings are still on e1 and e8, not castled. I have 15 white pieces, 16 black pieces.
. P . . . P p P	B P . . N . . .	5. Alice: Okay Bob. My board has fewer pieces than yours: 13 white pieces and 14 black pieces. Several pieces have been captured. Yours must be
. . P . P . P .	 P . . P	6. Bob: Okay Alice. My board has more pieces and pawns than yours. My Kings haven't moved or castled yet. Yours have. Therefore,
P . . P	P . P P N P P .	7. Alice: Okay Bob. Your board has more pieces and pawns than mine. Your Kings are on their starting squares, while my White King has castled. This confirms
. R B Q . K N .	R . B Q K . . n	8. Bob: _MINE_

Figure 2: Successful chess dialogue from Gemini 2.5 Pro with thinking. Chess is the sole task in which the models regularly exceed the token budget, as seen in turns 1, 5, 6, and 7.

private information X_i to a string of length $\leq n$. Then for any (n, c) , a game is level- k interactive if there exist encoding functions $f_1 \in \mathcal{F}_1^n, f_2 \in \mathcal{F}_2^n$ such that by application of these encoding functions over k turns of dialogue, player 1 can achieve an expected score $> c$. For example, a game is level-0 interactive if player 1 can answer without player 2's help; a game is level-1 interactive if there is an efficient encoding of X_2 such that the game is level-0 interactive; and so on. Clearly, games that are level- k interactive are also level- $(k+1)$ interactive; **we seek games such that for some $k \geq 2$, the game is level- k interactive but not level- $(k-1)$ interactive**, for reasonable values of n and c . A related property is the game should be solvable by a single player with access to both X_1 and X_2 . For a more formal discussion of levels of interactivity, with examples, see Appendix B.

3 Private information games

The MT-PINGEVAL tasks aim for high levels of interactivity by giving each player private information that is difficult to transmit efficiently as text. The games are chosen to cover a few distinct forms of private data and modes of interaction, including images and structured data. Some games are perfectly symmetric between the players (chess, COVR, name-game), while others assign distinct roles (MD3 and tangram). We now describe each game in detail. Full prompts are shown in Appendix D.

Chess. In the chess task, each player receives a board configuration in ASCII form. The boards are from the same game, and the players must determine which board is earlier in the game. The boards are chosen to be several moves apart, so that this should almost always be unambiguous. All moves are selected randomly from the uniform distribution over legal moves at each board position. Either player can end the game by saying _YOURS_ or _MINE_. An example is shown in Figure 2.

The task has a baseline success rate of 50%. An effective strategy would be to focus on two aggregate statistics: the number of pieces on the board and the rank of the most advanced pawns. Because these statistics are almost always monotonic over the course of a chess game, they could quickly identify the earlier board in most cases. The strategy of counting pieces is used by Gemini 2.5 Pro (see Figure 2), but is not discovered by the other models that we tested (Gemini 2.5 Flash, GPT-4o, Qwen-VL8B, and Gemma3-12B).

COVR. Multimodal reasoning benchmarks such as NLVR (Suhr et al., 2017) and COVR (Bogin et al., 2021) require answering natural questions or evaluating sentences about sets of images. These benchmarks can easily be adapted to conversational evaluation by showing one image to each player and requiring them to communicate about their images so as to arrive at an answer. We use data from the COVR benchmark, restricting ourselves to the examples with exactly two images; the generalization to multi-party conversations about larger numbers of images is left for future work. An example is shown in Figure 3. As shown in the prompt (Table D.2), it was necessary to instruct the language models on the desired answer format to enable exact match evaluation.

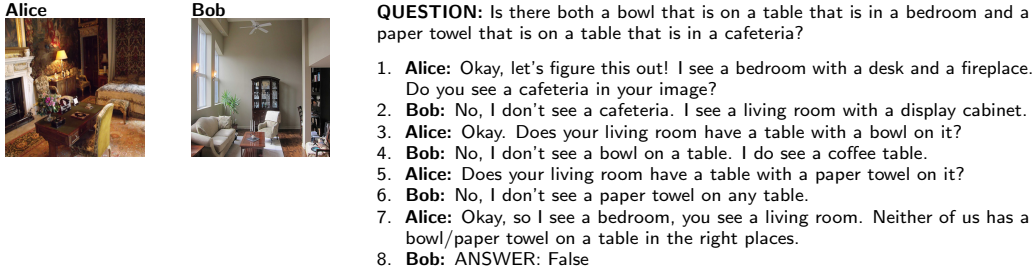


Figure 3: An example of a successful COVR dialogue from Gemma3-12B.

Image selection: MD3 and Tangram. The MD3 and tangram tasks are asymmetric image selection tasks. One player (the “describer”) receives a single image and the other (the “guesser”) receives a set of six photographs (MD3) or four geometrical shapes (tangram). The guesser must identify which of its images is the one seen by the describer, or to indicate if there is no match. This can be done by the describer enumerating features of their image, such as the focal object, the background, and visual features of the image’s components. The guesser can ask questions that distinguish likely candidates. These tasks are derived from prior work on descriptions of geometric objects (Ji et al., 2022) and human-human dialogue (Eisenstein et al., 2023). Examples of MD3 and tangram are shown in Figure 4.

Name-game. In the name-game task, each player receives a structured database of people, and the task is to find the single record that is common to both databases. To modulate difficulty, the size of the database and number of features can be changed; we include examples with 9, 16, and 25 records and five features. This task is inspired by He et al. (2017), although we generated new data using Gemini 2.5 Pro. An example of this task is shown in Figure 5. The task is quite challenging, and informally we have found that humans have a hard time identifying a good strategy.

4 Isotoken multi-turn scaling evaluation

To evaluate the multi-turn scaling capabilities of language models, we propose a new form of evaluation, called *isotoken evaluation*, which is motivated by the theory of levels of interactivity that was introduced in Section 2. The core idea is that both players are given a fixed total token budget, and we vary the number of turns into which that budget must be partitioned. For example, a budget of 128 tokens per player (256 tokens in total) may be partitioned into 16 tokens per turn in a 16-turn setting, or 128 tokens per turn in a 2-turn setting. The intuition is that any dialogue that can be performed in t turns can also be performed in $> t$ turns, because the players can ignore some of the turn boundaries. Thus, the task success should never decrease as the number of turns is increased, and for highly interactive tasks, we should expect task success to improve as the turn budget grows.¹

Both the token and turn budgets are upper limits: we do not force the models to use all available turns, as part of the challenge is to determine when the task has been successfully solved. For simplicity of implementation, we do not allow the models to partition the token budget arbitrarily, but rather allow exactly $\lfloor T/t \rfloor$ tokens per turn with a budget of T tokens and t turns. The token limit is given in the prompt, and all models were able to consistently stay under budget, except in the chess task.² In the rare cases where the token budget was violated, we truncate the output. The models sometimes disregarded the turn budget,

¹For comprehensive evaluations across all combinations of token and turn budgets, see Appendix C.

²The token budget is implemented in subword tokens, but the language models appear to be capable of reasoning only about whitespace-delimited tokens: for example, when in thinking mode, Gemini 2.5 Pro and Flash explicitly count out the number of whitespace-delimited tokens in each proposed response before responding. As there are approximately 1.3 subword tokens per whitespace-delimited token, we have the prompt state a token budget that is $1.3\times$ smaller than the true subword

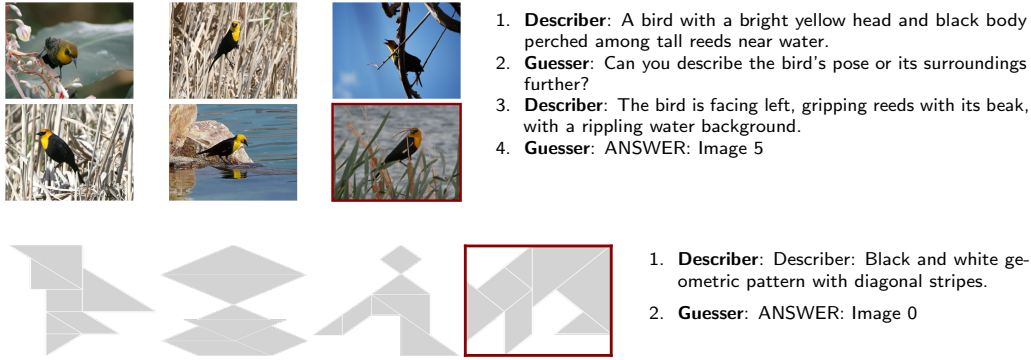


Figure 4: Top: example of successful dialogue for GPT-4o on the MD3 image selection task; the guesser sees all six images, while the describer sees only the highlighted image on the lower right. Bottom: unsuccessful dialogue for Qwen-VL8B on the Tangram task; the guesser sees all four images, while the describer sees the highlighted image on the far right.

failing to complete the task on the final turn. This problem was cured by adding reminders in the prompts for the final and penultimate turns, as shown in the bottom of Table D.3.

We apply isotoken evaluation to the MT-PINGEVAL tasks described in Section 3, with budgets of $\{2, 4, 8, 16\}$ turns and 128 tokens per player. We generate 1000 examples per task and evaluate five models: Gemini 2.5 Pro, Gemini 2.5 Flash, GPT-4o, Qwen-vl8B, and Gemma3-12B. The Gemini models are evaluated with and without thinking. (Thinking does not count against the token budget because the goal is a comparison that is *communication-matched*, not *compute-matched*.) All models use the same prompts, shown in Appendix D.

Summary of results. The main results are shown in Figure 1. With few exceptions, performance does not improve as the models are allowed more turns. In the image selection tasks, performance clearly gets worse as the turn budget increases. Figure 7 applies this evaluation to newer Gemini models, which show similar trends. We now describe some qualitative differences in the scaling trends across tasks, along with discussion of how the models approach the tasks.

Chess. In the chess task, the thinking models are clearly superior, and the non-thinking Gemini models improve somewhat with the number of turns. The remaining models perform at around the chance level of 50%. As noted above, on this task it is possible to give a highly informative summary of a board position in a very small number of tokens, by listing the number of pieces per player and the rank of the most advanced pawns (with few exceptions, these indicators move monotonically over the course of a game). This is shown in Figure 2, where both models report on the number of pieces on the board. Gemini 2.5 Pro is the only model to use this strategy, and it sometimes makes mistakes in counting. Other models tend to list apparently random subsets of their board positions before summarizing the situation and guessing at a conclusion (e.g., Figure E.13).

COVR. On the COVR task, which focuses on image analysis, the accuracies for all models are relatively flat as the number of turns increases. For the strongest model, Gemini 2.5 Pro (thinking), performance decreases monotonically from an overall best of 61.5% accuracy at two turns, to 58.8% at 16 turns. However, the non-thinking version of the same model improves its performance from 53.2% to 59.3% from two to 16 turns, suggesting that thinking and interactive dialogue may in some cases play similar roles. Most of the other differences in performance across turn budgets are small and statistically insignificant.

Image selection: MD3 and Tangram. The image selection tasks show the strongest evidence of inverse scaling with the number of turns: all models suffer from reduced performance as the number of turns increases and the number of tokens per turn decreases. Any 2-turn

token budget. This may partly explain the difficulty in the chess task, where the number of subwords per whitespace token would be higher for chess-specific board position terminology like ‘A8’.

Alice						Bob					
row	name	astrological sign	company	favorite musician	allergies	row	name	astrological sign	company	favorite musician	allergies
1	Michael	Scorpio	Cisco	John Coltrane	Corn	1	Isabella	Leo	Tesla	Miles Davis	Eggs
2	Andrew	Aries	Amazon	Count Basie	Shellfish	2	Lily	Sagittarius	HP	Herbie Hancock	Soy
3	Ethan	Cancer	Tesla	Miles Davis	Eggs	3	Chloe	Pisces	Oracle	Herbie Hancock	Mustard
4	Asher	Aries	Sony	Ella Fitzgerald	Corn	4	Chloe	Gemini	Qualcomm	Louis Armstrong	Gelatin
5	Lucas	Capricorn	Amazon	Art Blakey	Fish	5	Sophia	Libra	Samsung	Wes Montgomery	Peanuts
6	Charlotte	Libra	Cisco	Nina Simone	Mustard	6	Aurora	Capricorn	Apple	Billie Holiday	Sesame
7	Isabella	Leo	Tesla	Miles Davis	Eggs	7	Sofia	Leo	Amazon	Art Blakey	Tree Nuts
8	Asher	Cancer	Google	Ella Fitzgerald	Corn	8	Ella	Pisces	Apple	Thelonious Monk	Tree Nuts
9	Isabella	Leo	HP	Miles Davis	Eggs	9	Samuel	Leo	Tesla	Miles Davis	Corn

1. **Alice:** Hello! I know a Michael who works at Cisco. Do you know him?

2. **Bob:** I don't know a Michael at Cisco. Do you know anyone who works at Apple?

3. **Alice:** I don't know anyone at Apple. I know an Andrew who works at Amazon. Do you know him?

4. **Bob:** I don't know Andrew. Do you know anyone who is a Leo?

5. **Alice:** Yes, I know two Leos. One works at Tesla. Do you know an Ethan who works at Tesla and likes Miles Davis?

6. **Bob:** I don't know Ethan. I know two Leos at Tesla liking Miles Davis, but different names. Do you know a Pisces?

7. **Alice:** I don't know a Pisces. I know one Leo at Tesla liking Miles Davis. Her name is Isabella. Do you know her?

8. **Bob:** SELECT ROW 1

Figure 5: A successful but lucky dialogue for Gemini 2.5 Flash (thinking) on the name-game.

dialogue could in principle be played over a larger number of turns, so there is nothing that makes the many-turn / few-token setting inherently more challenging. Some models waste tokens in each turn with uninformative introductions, but the main issue is that they do not use their full turn budgets, to a far greater extent than in the other tasks (Figure 8). For example, on MD3, even Gemini 2.5 Pro uses fewer than four turns on average in the setting where it is permitted 16 turns and 16 tokens per turn. As a result, the total number of tokens used by each player decreases as the number of turns grows (Figure 8). This is despite our efforts to tune the prompts to encourage the models not to terminate the dialogue until completely sure of their answer (Table D.4). Premature termination has been identified as a cause of failures in practical multi-LLM applications (Cemri et al., 2025).

Name-game. Name-game stands out as the task in which performance improves substantially with the number of turns. However, inspection of the transcripts shows that in many cases, the players simply guess randomly until they find a lucky match or run out of turns. This can be seen in Figure 5, where Alice happens to mention Isabella in the final turn, and Bob marks this as a match without confirming all of the attributes. The example also highlights the questionable pragmatics of even relatively strong models: Alice’s response in turn 5 strongly implies that Ethan is the astrological Leo who works at Tesla and likes Miles Davis, but in fact she has changed the topic and is no longer talking about Leos.

While the language models do improve their performance with the turn budget, this is driven primarily by poor performance at low turn budgets rather than effective use of interaction at high turn budgets. The models usually guess only a single row per turn, and therefore do not use all allowed tokens in the 2-turn and 4-turn cases. Indeed, a baseline that simply proposes a random record in each turn (sampled without replacement) displays a similar scaling trend, achieving a higher success rate than all models except Gemini 2.5 Pro (thinking) in the 16-turn setting (Figure 6). By the nature of the task, this baseline’s performance will improve rapidly with the turn limit. This explains much of the improvement in performance of the language models, which follow a similar strategy.

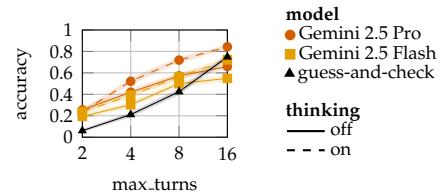


Figure 6: Comparison of the Gemini models against a simple guess-and-check heuristic on name-game.

5 Linguistic analysis

To better understand the performance of language models on MT-PINGEVAL, we explore three possible explanations in the linguistic features of model output:

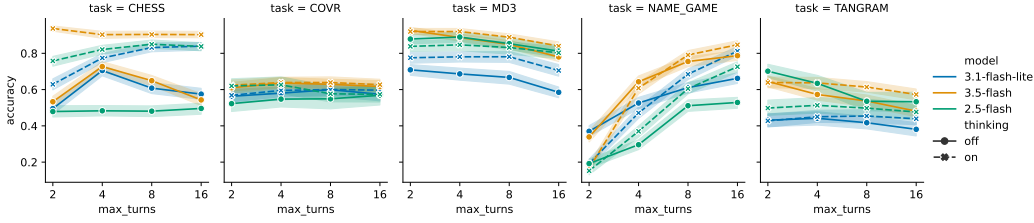


Figure 7: Isotoken evaluation of Gemini Flash models

- **sycophancy** could cause LLMs to disregard their own private information, leading to collaborative failure when their private information is in fact crucial;
- low **information density** might cause LLMs to struggle under token constraints;
- lack of **goal-directedness** might cause LLMs to fail to elicit necessary information from the other player.

We first present linguistic analyses of the model outputs (§5.1), and then offer a rough comparison with similar human data from the MD3 task (§5.2). We find strong evidence against the first two hypotheses: LLMs are in fact quite capable of rejecting proposals that are incompatible with their private information (fig. 9), and LLM dialogues are significantly more information-dense than comparable human dialogues (fig. 10). Goal directedness is not definitively excluded as an explanation, but it is suggestive that the most successful models (Gemini thinking models) are not the ones with the highest coherence.

5.1 Features

Sycophancy. Sycophancy is a major concern about today’s large language models (Perez et al., 2023), due in part to its potential effects on human-human interaction patterns (Cheng et al., 2025). In the context of MT-PINGEVAL, sycophantic language model players might be too willing to accept claims or ideas by the other player, potentially overriding private information that could help them achieve a more accurate answer (Stengel-Eskin et al., 2025). If so, this would suggest that language models are likely to do the same in dialogues with human interlocutors, which would significantly limit their utility. We tested for sycophancy in two ways: (1) focused analysis of how models handle proposed answer hypotheses in the chess task (where answers are binary and the player roles are symmetric), and (2) use of LLM-as-a-judge to detect uncritical agreement and validation of false premises. We found that while there is considerable evidence for “stylistic” sycophancy according to the LLM judge, the language models evaluate each other’s answer hypotheses critically, and are far more likely to accept correct hypotheses. For more details, please see Appendix E.1 and E.2.

Information density. The isotoken evaluation paradigm leans heavily on the capability of models to communicate *efficiently* — with high information density — due to the evaluation’s use of tight token constraints. On the other hand, more interactive dialogue naturally requires conversational scaffolding (e.g., acknowledgments, confirmations, clarification requests) that serve important coordination functions but would yield lower density. A natural question, then, is whether the models that achieve the greatest task success are those that maintain high levels of information density even in settings with many turns and few tokens per turn. To test this, we focus on *lexical density* (Ure, 1971; Halliday, 1985), a widely used measure in linguistics and discourse analysis that captures how much informational content a text conveys by examining the proportion of lexical (content) words relative to the total number of words. We implement a lexical density metric that combines two complementary signals: (1) the *content-word ratio*, which measures the proportion of content that excludes fillers, stopwords, and disfluencies; (2) the *novelty* of each turn, as measured by TF-IDF (Jones, 1972), which is widely used to measure informativeness of words in a corpus. We define lexical density as the product of these two quantities. However, we find limited support for the hypothesis that stronger models would maintain higher density. Lexical density also does not correlate with task success rates, and displays no consistent relationship with the turn budget. **This discrepancy between communicative effort and**

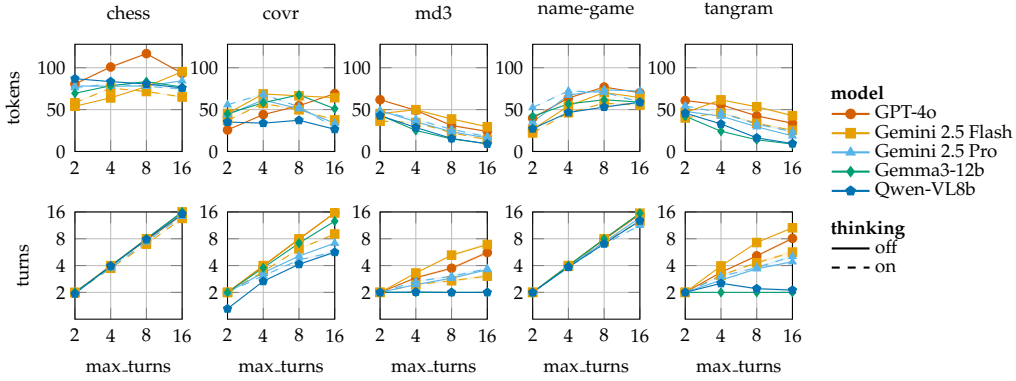


Figure 8: Tokens and turns per dialogue.

task success suggests that current LLMs can generate content-rich utterances but struggle to deploy that content strategically in service of collaborative goals. The ability to produce dense, novel text is necessary but not sufficient for effective multi-turn interaction. For more details, please see Appendix E.3

Goal directedness and coherence. Goal-directed conversations are coherent in the sense of maintaining focus on the overarching goal of interaction. Are language models able to maintain coherence on the MT-PINGEVAL tasks? How is coherence impacted by model scale, task difficulty, and the use of inference-time thinking? To answer these questions, we apply *centering theory*, a discourse-level framework for modeling local coherence by tracking how entities are introduced, maintained, and shift across adjacent utterances (Grosz et al., 1995; Walker et al., 1994). Historically, centering theory has played a number of roles; for our purpose, most critical is its role as a foundation for modeling coherence in text generation and evaluation, providing interpretable signals that correlate with human judgments of discourse quality (Barzilay & Lapata, 2008; Miltsakaki & Kukich, 2000). The key property of centering theory is that it focuses on *local transitions*, capturing how speakers manage continuity and change in what a discourse is “about” from one utterance to the next. This is implemented through the notion of *centers*, discourse entities that are salient at a given point; more coherent discourses are more likely to feature utterance-to-utterance transitions that retain the same discourse elements in the central positions. We operationalize these principles into a heuristic measure of centering-based coherence, by tracking the centering transitions between adjacent utterances; for details on the method, please see Appendix E.4. We find that in general among the non-thinking models, **the most powerful and effective models also produce the most coherent dialogues; however, surprisingly, thinking significantly reduces coherence** even as it improves performance on some of the MT-PINGEVAL tasks. This may be because our lexico-syntactic measure of coherence combines multiple distinct phenomena. In some cases, the coherence score reflects the desired goal-directed focus of the participants. In other cases, it reflects relatively superficial features, such as when a turn ends with a question and the next turn begins with an explicit answer to that question, e.g. “Do you know a Pisces? / I don’t know a Pisces.”

5.2 Comparison with human dialogues

A closely-related image selection task is performed by humans in prior work (Eisenstein et al., 2023).³ The image sets are the same, but there are a few important differences. The humans had no limit on the turns or tokens that they could use, but were asked to solve as many task instances as they could in five minutes. “No match” was not an option, but the number of distractor images was 12 rather than six. With those caveats in mind, it is possible to make a rough comparison between the properties of human-human dialogues with those of dialogues between language models (Shaikh et al., 2024). While this comparison is imperfect, it represents a first step beyond prior work on evaluation of LLM multi-turn

³These transcripts are available at <https://www.kaggle.com/datasets/jacobeis99/md3en>

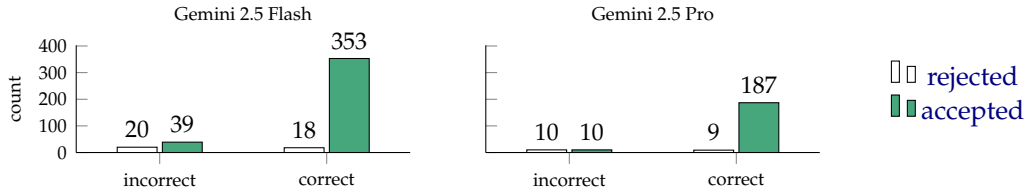


Figure 9: Counts of early answer proposals in the chess task for the Gemini 2.5 thinking models. Some incorrect proposals are accepted, but at a far lower rate than correct proposals.

capabilities, which makes no comparison with human language at all (e.g., Bai et al., 2024; Kwan et al., 2024; Laban et al., 2025).

Task performance. Humans achieve task success rates between 86-91%. This exceeds the task success rates for language models, although the success rates are not directly comparable for the reasons noted above.

Efficiency. Humans averaged 10.6 whitespace-delimited tokens per turn (median = 5), and 5.5 turns per prompt (median = 4). This means that on average humans used only 60 tokens per dialogue in total (across both players). The short turn lengths and relatively interactive dialogues may be linked to the speech modality (Biber, 1991). While the language models were *allowed* 256 tokens per dialogue, they *used* significantly fewer: as shown in Figure 8, each player used fewer than 50 tokens in all settings, except GPT-4o at 2 turns. Thus, the language models were less token-efficient than the human players, while also failing to fully utilize their token budget to maximize task performance.

Linguistic features. Human dialogues had an average lexical density of 24.95; this was lower than all of the language models on all versions of MD3, which ranged from 40–70. Again, a likely explanation is the speech modality: 89% of the human dialogues contained a backchannel or filler token (“ok”, “okay”, “done”, “um”, “uh”) and 6.8% contain an apology. The human dialogues have an average coherence score of 0.86, which exceeded most of the models in this task, except for Gemini 2.5 Pro (no thinking). The relationship between lexical density and coherence on this task is shown in Figure 10.

In summary, humans are able to perform the MD3 image selection task with high accuracy, despite low lexical density. This may be partly explained by the relatively high coherence of human dialogues, which exceeds all but one of the language models. Beyond coherence, the outstanding token efficiency of human dialogues suggests that they approach these tasks in a more goal-directed way than even the strongest language models.

6 Related work

Early multi-turn evaluations of large language models focused on manually-designed follow-up questions to elicit multi-turn conversations (Zheng et al., 2023; Kwan et al., 2024; Bai et al., 2024). Laban et al. (2025) describe these as “episodic” evaluations, because each turn introduces a subtask that can in principle be evaluated in isolation. They propose an alternative evaluation in which single-turn questions are sharded into multiple turns, which they find to significantly diminish the performance of even the strong language models. However, the sharding is an arbitrary source of noise rather than the natural result of attempts to communicate private information under information bandwidth constraints.

Self-play has been widely used to generate training and evaluation data for task-oriented dialogue, typically involving a user simulator and an assistant (Shah et al., 2018; Wang et al., 2023; Li et al., 2024; Shao et al., 2024; Dou et al., 2025). Although this setup mirrors deployment, it is challenging for evaluation of the multi-turn collaboratively capability because user simulation requires faithfully reproducing the many foibles of human speakers, such as laziness, imprecision, and reliance on untested assumptions (Lin & Tomlin, 2025). Progress on this orthogonal capability of user simulation may alter multi-turn performance in unpredictable ways. As simulators improve, performance could become worse, because

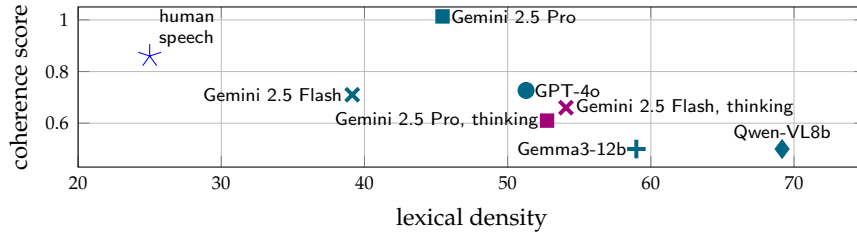


Figure 10: Lexical density and coherence for AI models and humans on the MD3 image selection task. The language model scores are averaged over all turn budgets.

the simulators more faithfully imitate grounding failures and comprehension gaps. Alternatively, performance may appear to improve, because stronger user simulators may structure their communication more coherently. Private information games sidestep these issues: the language models do not need to simulate human users and are free to make their best efforts to achieve task success under the given constraints.

Two papers take a similarly symmetrical approach to evaluation of collaboration between language model agents, finding significant challenges in multi-agent collaboration. [Davidson et al. \(2025\)](#) design a navigation task in which each agent sees a partially-occluded version of the same map, and must communicate to identify a path. They find that this setting is significantly more challenging than one in which a single agent sees the whole map. In CooperBench, a set of coding tasks is divided between two agents ([Khatua et al., 2026](#)), showing that agents fail to coordinate their work (see also [Cemri et al., 2025](#)). Our approach focuses on simpler tasks that can be completed with a single, short output. This allows us to quantify the scaling of language model collaboration across varying turn budgets and to compare linguistic phenomena across diverse types of tasks.

Unconstrained self-play evaluation has also been applied in the chit-chat setting by [Ghan-deharioun et al. \(2019\)](#), who focus on measures of sentiment and semantic coherence across the dialogue (see also [Gooding & Grefenstette, 2025](#)). We explore measures of coherence in Appendix E.4, but rather than using word and sentence embeddings, we build on linguistic theories about the role of discourse entities through the conversation. Another distinction is that we consider dialogues in which the participants must share information to accomplish a task, rather than more open chitchat (see also [Duan et al., 2024](#), for a chitchat evaluation that focuses on whether LLM-generated turns can be detected in human-human dialogues).

7 Conclusion

We have presented MT-PINGEVAL, a benchmark for evaluating multi-turn conversational capabilities through private information games. The isotoken evaluation methodology fixes the total communication budget while varying the number of turns, isolating multi-turn interaction capabilities from raw task-solving ability. This reveals a consistent pattern: language models generally fail to improve, and often get worse, as they are given more opportunities for interactive communication. Linguistic analysis provides some partial explanation: models exhibit stylistic sycophancy (e.g., spurious apologies and uncritical agreement), wasting tokens without advancing the task. And despite generating locally coherent dialogues, this coherence reflects formulaic patterns rather than strategic progress toward shared understanding. Comparison with human performance on the image guessing task highlights the gap: humans achieve higher accuracy with far fewer tokens, suggesting the challenge lies not in capability but in communication strategy — knowing what to share, what to elicit, and when to commit.

Private information is a defining feature of real-world communication: we rarely know exactly what our interlocutors know, want, or need. MT-PINGEVAL offers a rigorous framework for measuring progress on this fundamental challenge, and our results suggest that substantial advances in multi-turn interaction remain ahead.

Acknowledgments Thanks to Diego Antognini and Kevin Robinson for their assistance with engineering aspects of the project. The paper benefited from feedback from Bernd Bohnet, Ryan Burnell, Elizabeth Clark, William Cohen, Chris Dyer, Asma Ghandeharioun, Anastasios Kement, Tal Linzen, Annie Louis, Claudiu Musat, and Hannah Rashkin. Thanks also to the CoLM reviewers for several constructive suggestions that improved the paper.

References

- Ali Ahmadvand, Jason Ingyu Choi, and Eugene Agichtein. Contextual dialogue act classification for open-domain conversational agents. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–42, 2019.
- Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*, 2024.
- Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008. doi: 10.1162/coli.2008.34.1.1. URL <https://aclanthology.org/J08-1001/>.
- Douglas Biber. *Variation across speech and writing*. Cambridge university press, 1991.
- Ben Bogin, Shivanshu Gupta, Matt Gardner, and Jonathan Berant. Covr: A test-bed for visually grounded compositional generalization with real images. *arXiv preprint arXiv:2109.10613*, 2021.
- Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pp. 155–162, 1987.
- Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, et al. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*, 2025.
- Haixia Chai and Michael Strube. Incorporating centering theory into neural coreference resolution. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2996–3002, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.218. URL <https://aclanthology.org/2022.naacl-main.218/>.
- Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sympathetic ai decreases prosocial intentions and promotes dependence. *arXiv preprint arXiv:2510.01395*, 2025.
- Tim R Davidson, Adam Fourney, Saleema Amershi, Robert West, Eric Horvitz, and Ece Kamar. The collaboration gap. *arXiv preprint arXiv:2511.02687*, 2025.
- Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. Simulatorarena: Are user simulators reliable proxies for multi-turn evaluation of ai assistants? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35200–35278, 2025.
- Catherine Doughty and Teresa Pica. “Information gap” tasks: Do they facilitate second language acquisition? *TESOL quarterly*, 20(2):305–325, 1986.
- Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. Botchat: Evaluating llms’ capabilities of having multi-turn dialogues. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 3184–3200, 2024.

- Jacob Eisenstein, Vinodkumar Prabhakaran, Clara Rivera, Dorottya Demszky, and Devyani Sharma. Md3: The multi-dialect dataset of dialogues. *arXiv preprint arXiv:2305.11355*, 2023.
- Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. Approximating interactive human evaluation with self-play for open-domain dialog systems. *Advances in Neural Information Processing Systems*, 32, 2019.
- Sian Gooding and Edward Grefenstette. Interaction dynamics as a reward signal for llms. *arXiv preprint arXiv:2511.08394*, 2025.
- Peter C. Gordon, Barbara J. Grosz, and Laura A. Gilliom. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17(3):311–347, 1993.
- H. P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics: Vol. 3: Speech Acts*, pp. 41–58. Academic Press, New York, 1975. URL <http://www.ucl.ac.uk/ls/studypacks/Grice-Logic.pdf>.
- Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.
- M. A. K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold, London, 1985.
- He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. *arXiv preprint arXiv:1704.07130*, 2017.
- Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 582–601, 2022.
- Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.
- Nikiforos Karamanis, Chris Mellish, Massimo Poesio, and Jon Oberlander. Evaluating centering for information ordering using corpora. *Computational Linguistics*, 35(1):29–46, March 2009. doi: 10.1162/coli.07-036-R2-06-22. URL <https://aclanthology.org/J09-1003/>.
- Arpandeeep Khatua, Hao Zhu, Peter Tran, Arya Prabhudesai, Frederic Sadrieh, Johann K Lieberwirth, Xinkai Yu, Yicheng Fu, Michael J Ryan, Jiaxin Pei, et al. Cooperbench: Why coding agents cannot be your teammates yet. *arXiv preprint arXiv:2601.13295*, 2026.
- Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*, 2024.
- Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.
- Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.
- Jessy Lin and Nick Tomlin. User simulators bridge RL with real-world interaction. <https://jessylin.com/2025/07/10/user-simulators-1/>, July 2025. Accessed: 2025-11-19.
- Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. Decision-oriented dialogue for human-ai collaboration. *Transactions of the Association for Computational Linguistics*, 12: 892–911, 2024.

- Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic Turing test: Learning to evaluate dialogue responses. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1116–1126, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1103. URL <https://aclanthology.org/P17-1103/>.
- Kazuki Matsumoto et al. Topic break detection in interview dialogues using BiLSTM-CRF. *Sensors*, 22(2):694, 2022.
- Eleni Miltsakaki and Karen Kukich. The role of centering theory’s rough-shift in the teaching and evaluation of writing skills. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 408–415, Hong Kong, October 2000. Association for Computational Linguistics. doi: 10.3115/1075218.1075270. URL <https://aclanthology.org/P00-1052/>.
- Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL <https://aclanthology.org/2023.findings-acl.847/>.
- Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363, 2004. doi: 10.1162/0891201041850911. URL <https://aclanthology.org/J04-3003/>.
- Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In Srinivas Bangalore, Jennifer Chu-Carroll, and Yunyao Li (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pp. 41–51, New Orleans - Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-3006. URL <https://aclanthology.org/N18-3006/>.
- Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6279–6296, 2024.
- Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. Collaborative Gym: A framework for enabling and evaluating human-agent collaboration, 2024.
- Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. Teaching models to balance resisting and accepting persuasion. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8108–8122, 2025.
- Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017.

- Jean Ure. Lexical density and register differentiation. In G. E. Perren and J. L. M. Trim (eds.), *Applications of Linguistics*, pp. 443–452. Cambridge University Press, Cambridge, 1971.
- Marilyn A. Walker, Masayo Iida, and Sharon Cote. Japanese discourse and the process of centering. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 88–95, Las Cruces, New Mexico, 1994. Association for Computational Linguistics.
- Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*, 2023.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.
- Xiaodan Zhu and Gerald Penn. Utterance-level extractive summarization of open-domain spontaneous conversations with rich features. In *2006 IEEE International Conference on Multimedia and Expo*, pp. 793–796, 2006. doi: 10.1109/ICME.2006.262600.

Contents

A	Language model usage	16
B	Levels of interactivity	16
C	Additional statistics	17
D	Private Information Game Prompts	17
E	Linguistic analysis: additional results	17
E.1	Sycophancy in task performance	21
E.2	Stylistic sycophancy	22
E.3	Information density	23
E.4	Goal directedness and coherence	25
F	Text analysis methods	28
F.1	Answer proposals	28
F.2	Stylistic sycophancy autorater prompt	29
F.3	Information density	29
F.4	Goal directedness and coherence	30

A Language model usage

We used language models in the following ways: (1) creating figures and tables by generating LaTeX code; (2) helping write data analysis code; (3) formatting bibtex entries; (4) automatically assessing sycophancy, as described in Appendix E.2; (5) generating the name-game examples, as described in Section 3; (6) reviewing the paper and suggesting areas in which the writing could be improved.

B Levels of interactivity

To formalize intuitions about the role of private information in multi-turn communication, we propose a framework for quantifying the *level of interactivity* for a collaborative private information game. Consider a two-player game, where player 1 receives private information X_1 and player 2 receives private information X_2 . The shared goal is for player 1 to give an answer that gets a high score. (For this theoretical formulation it is convenient to allow only player 1 to give an answer.) The private information need not be completely disjoint: for example, both X_1 and X_2 may contain some description of the task, but they should also each contain at least some information that is unique.

Assume each player $i \in \{1, 2\}$ has an encoding function $f_i \in \mathcal{F}_i^n = \{f : \mathcal{X}_i \times \mathcal{S} \rightarrow \mathcal{V}^n\}$, where $\mathcal{V}^n$ is the set of strings with n tokens on vocabulary $\mathcal{V}$ and $\mathcal{S} = (\mathcal{V}^n)^*$ is the set of possible conversational histories. The history is then updated by concatenation, $S_t := S_{t+1} \circ f_{i(t)}(X_{i(t)}, S_{t+1})$, with t indicating the number of remaining turns and $i(t)$ indicating the active player at turn t . We will use $i(t) = t \bmod 2 + 1$, with $i(0) = 1$ indicating that player 1 always provides the final answer. At the initial turn T , we have $S_T = \emptyset$, and we will write $f(X)$ as shorthand for $f(X, S_T = \emptyset)$. Finally, let $\mathbb{E}[\text{SCORE}(X_1, S)] \geq 0$ represent the expected score from player 1 answering based on private information X_1 and state S . We give concrete examples of such games and encoding functions immediately below.

For constants $n \in \mathbb{Z}_+$, $c \in \mathbb{R}_+$, we can characterize games by their level of interactivity:

- Level 0: The game is solvable without X_2 , so no communication is needed: $\mathbb{E}[\text{SCORE}(X_1, \emptyset)] > c$.
- Level 1: There is an efficient encoding function for X_2 that makes the game level-0 interactive:

$$\exists f_2 \in \mathcal{F}_2^n : \mathbb{E}[\text{SCORE}(X_1, S_0)] > c$$

where $S_0 = f_2(X_2)$.

This means that the game is solvable in one round of conversation, with player 2 transmitting the length-restricted message $f_2(X_2)$ to player 1. For example, suppose that the game is image classification, so that X_2 is an image and Y labels the class of the main object in the image. Player 2 does not need to know anything about X_1 (indeed, X_1 can even be empty) to transmit the message $Y = f_2(X_2)$ to player 1.

- Level 2: There is an efficient encoding function for X_1 such that sharing this encoding makes the game level-1 interactive:

$$\exists f_1 \in \mathcal{F}_1^n, \exists f_2 \in \mathcal{F}_2^n : \mathbb{E}[\text{SCORE}(X_1, S_0)] > c$$

where $S_1 = f_1(X_1)$
 $S_0 = S_1 \circ f_2(X_2, S_1) = f_1(X_1) \circ f_2(X_2, f_1(X_1))$

For intuition, suppose that X_2 is an image and X_1 is another image and a text proposition (e.g., “there are more beads in my image than your image”, $f_1(X_1) =$ “how many beads”, $f_2(X_2, f_1(X_1)) =$ “three”), and Y labels whether the proposition is true. This example corresponds to our multiplayer implementation of COVR (Bogin et al., 2021), detailed in Section 3, although some instances from this task may not be resolvable in such a minimal dialogue.

- Level 3: There is an efficient encoding of X_2 such that sharing this encoding makes the game level-2 interactive:

$$\begin{aligned} \exists f_1 \in \mathcal{F}_1^n, \exists f_2 \in \mathcal{F}_2^n : \mathbb{E}[\text{SCORE}(X_1, S_0)] > c \\ \text{where } S_2 = f_2(X_2) \\ S_1 = S_2 \circ f_1(X_1, S_2) = f_2(X_2) \circ f_1(X_1, f_2(X_2)) \\ S_0 = S_1 \circ f_2(X_2, S_1) \\ = f_2(X_2) \circ f_1(X_1, f_2(X_2)) \circ f_2(X_2, f_2(X_2) \circ f_1(X_1, f_2(X_2))) \end{aligned}$$

It is harder to construct intuitive examples at this level and above. However, imagine the game of identifying elements that are shared between structured objects X_1 and X_2 . The game can be solved in the following fashion: $f_2(X_2)$ lists at a high level the elements present in X_2 ; $f_1(X_1, S_2)$ identifies a subset of those elements that are also present in X_1 ; $f_2(X_2, S_1)$ describes the detailed *features* of those elements in X_2 ; and given this information, player 1 outputs a list of elements that have identical features in both X_1 and X_2 .

- Level $k > 3$: There is an efficient encoding of X_1 or X_2 (depending on whether k is even or odd) such that sharing this encoding makes the game level $k - 1$ interactive.

Clearly games that are level- k interactive are also level $(k + 1)$ -interactive. We seek games such that for some $k \geq 2$, the game is level- k interactive but not level $(k - 1)$ -interactive, for reasonable values of n and c . A related property is the game should be solvable by a single agent with access to both X_1 and X_2 . Note that interactivity is a property of a game *and model*, because a strong model might be able to solve a game in a smaller number of turns than a weaker model by producing better encodings or making better inferences. It is therefore not possible to analytically compute a game’s level of interactivity. Rather, levels of interactivity are a conceptual tool to help formalize the design of our multi-turn evaluations.

C Additional statistics

In the isotoken evaluations, we hold the total token budget constant. We now present supplementary results for all combinations of turn and token budgets, for Gemini Flash versions 2.5 (fig. C.11) and 3.5 (fig. C.12). The results are grouped by tokens per turn. At fixed token-per-turn budgets, increasing the turn budget increases the overall communication bandwidth. Even in this setting, increasing the turn budget does not yield consistent performance improvements in the MD3, TANGRAM, and COVR tasks. As before, performance increases with the number of turns in NAME-GAME, due to the “guess-one-row-per-turn” strategy described in Section 4. In CHESS, Gemini 3.5 Flash demonstrates improvements with the turn budget, but for Gemini 2.5 Flash this is contingent on the thinking strategy: without thinking, there are some performance improvements when the models are allowed a very large overall token budget (≥ 1024 , which is 4x larger than the isotoken setting); with thinking, there are improvements over the 2-turn setting. We believe this is due to the increase in the total number of available tokens (constant tokens-per-turn, increasing number of turns), which allows the players to describe their boards more comprehensively.

D Private Information Game Prompts

Prompt examples for all the MT-PINGEVAL tasks are shown in Tables D.1, D.2, D.3, and D.4.

E Linguistic analysis: additional results

This gives more detailed results from the linguistic analysis. Methods are described in more detail in appendix F.

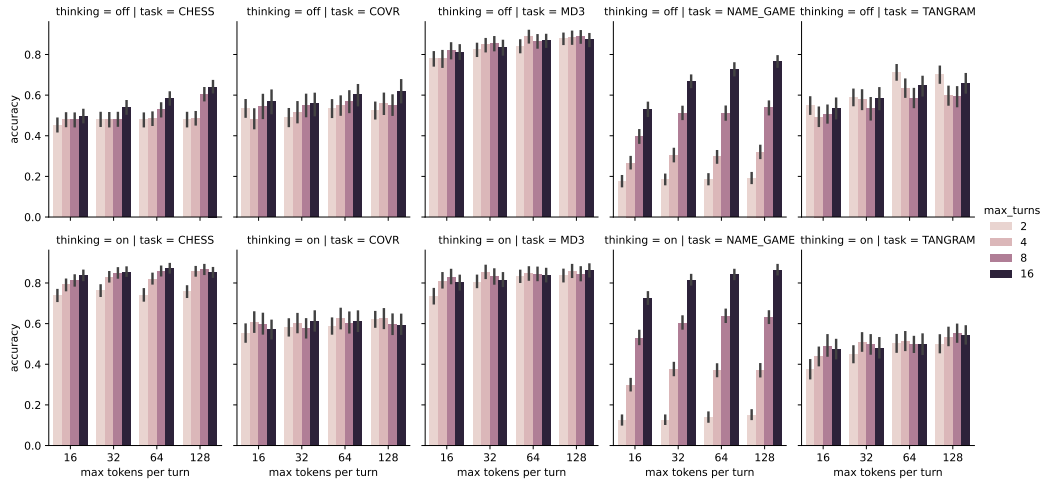


Figure C.11: Full turn/token budget sweep for Gemini 2.5 Flash

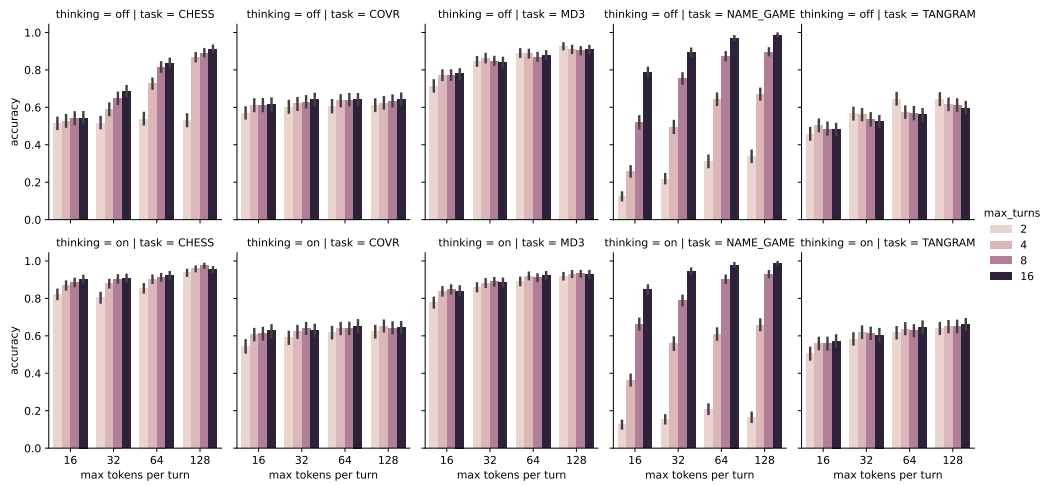


Figure C.12: Full turn/token budget sweep for Gemini 3.5 Flash

Your name is Bob and my name is Alice. You are a chess expert. We are going to talk about chess games using diagrams in which the white pieces are shown by upper-case letters like 'K' and 'Q', and the black pieces are shown by lower-case letters like 'k' and 'q'. Empty squares are shown as periods, '.'.

For example, in this board, White has moved its pawn to D4 and its knight to A3, while Black has moved its pawn to C5:

```
r n b q k b n r
p p . p p p p
. . . . .
. . p . . . .
. . . P . . .
N . . . . .
P P P . P P P
R . B Q K B N R
```

Each of us will be shown a board that the other player cannot see. The two boards are from the same game of chess. Our task is to determine which board came earlier in the game. To accomplish this, we must describe our boards to each other.

You can use only 22 words per turn, so you cannot fully specify the locations of all pieces. Instead, describe the key features that will help us determine which board came first.

After discussing it, if you think your board came first, say '_MINE_'; if you think my board came first, say '_YOURS_'. Otherwise, you can continue to describe your board to me, or ask questions about my board.

It is important to get this right, so let's discuss until we are sure. However, if I say this is your final turn, then you **must** give a final answer that is either '_MINE_' or '_YOURS_'.

Here is your board, which I cannot see:

```
. n . . q . n r
r b . p k . b .
. . P . . . .
p . . . P p p p
. . B P . . . .
. P . . . P P
P . P N K P . .
R N . Q . R . .
```

CONVERSATION

Here is our conversation history.

1. ME: Hello Bob. On my board, all black pieces are on their starting squares. White has moved the d-pawn to d4. No other white pieces

You have 7 turns left.

2. YOU:

Table D.1: Example chess prompt, after turn 1 of the dialogue.

INSTRUCTIONS

You are Alice, and you are playing a collaborative game that involves having a conversation with another player about two images. You can see the 'left' image and the other player can see the 'right' image.

Here is the 'left' image: <start_of_image>

With the help of the other player, you have to answer the question:
"How many images contain at least 1 women that are watching child that is wearing helmet?"

To do this, you will have a conversation with the other player, for example by sharing information about your image or asking about the other player's image. In each turn you can use at most 22 words, so be concise!

When you are ready, output 'ANSWER:' followed by your answer to the question. Your answer should be a *single* word, a number, or (rarely) a short phrase, e.g.:

```
`ANSWER: 4'
`ANSWER: True'
`ANSWER: next door'
```

If it is a true/false question, make sure to output 'True' or 'False'.

Table D.2: Example COVR prompt, on the first turn of the dialogue.

INSTRUCTIONS

Your name is Bob, and we are going to play a conversational game. You have a database of people who you know:

'''

row,name,company,favorite color,favorite musician,city

1,Noah,IBM,Mustard Yellow,Wes Montgomery,Chicago

2,Joshua,Samsung,Charcoal,Miles Davis,Sydney

3,James,HP,Charcoal,Charlie Parker,Rome

4,Madison,Google,Mustard Yellow,Thelonious Monk,Berlin

5,Elizabeth,Foxconn,Coral,John Coltrane,Rio de Janeiro

6,Asher,Amazon,Teal,John Coltrane,Moscow

7,Asher,Foxconn,Teal,John Coltrane,Moscow

8,Owen,Sony,Peach,Wes Montgomery,London

9,Madison,Foxconn,Coral,John Coltrane,New York

'''

I might know one of the same people. Let's talk to find out whom we both know. You can ask questions or describe some of the people you know. If I ask you a question, don't forget to answer it.

To match, *all attributes* of the row must match exactly.

If you are sure you have identified a person that we both know, output only `SELECT ROW i`, with `i` indicating the row number of your mutual friend. Only answer when you are completely sure, and talk it over until then. You can use only 44 words per turn, so be concise.

If I say that this is your final turn, then you **must** give a final answer in the format of `SELECT ROW i`.

CONVERSATION

Here is our conversation history.

1. ME: I know Chloe from Microsoft who loves crimson and Charles Mingus, based in Paris. Do you know her? Or someone else with matching details?

2. YOU: I don't know Chloe. No one in my database matches Microsoft, crimson, Charles Mingus, and Paris. Do you know anyone else with those traits?

3. ME: I know David from Sony, mustard yellow, Art Blakey, Berlin. Do you know him? Or any other with exact same traits?

This is your final turn, so you must give a final answer. Based on the conversation, make a guess and output `SELECT ROW i`. Don't say anything else: do not explain or ask any questions. Your next words **must** be `SELECT ROW` and your last word must be a number between 1 and 9.

Table D.3: Example name-game prompt, on the fourth and final turn of the dialogue. Note the extra instructions to finish the dialogue.

INSTRUCTIONS

We are playing an interactive game, which involves having a conversation. In this game you are the GUESSER and I am the DESCRIBER. I will show you a numbered list of images. Next, I will describe an image, which may be in your list. If you are sure that the image I am describing is in our list, respond with the number of the image I am describing.

For example, to identify image 1, your output should be formatted as 'ANSWER: Image 1' and nothing else.

Sometimes the image I am describing will not be in your list. If you are sure that none of the images in your list match mine, respond with 'ANSWER: No match' and nothing else.

If you are not sure which image matches the description, you can ask me a clarification question or tell me about your choices. In each turn you can use no more than 11 words, so be concise.

It's better to ask for clarification than to guess incorrectly, so don't be afraid to ask me for more information if you are not sure!

Images

Image 0: <start_of_image>
Image 1: <start_of_image>
Image 2: <start_of_image>
Image 3: <start_of_image>

CONVERSATION

Here is our conversation history.

1. ME: Black and white abstract pattern with diagonal stripes and geometric shapes.
2. YOU: Is the pattern mostly diagonal or horizontal?
3. ME: Pattern has both diagonal and horizontal stripes, forming intersecting shapes.

You have 13 turns left.

4. YOU:

Table D.4: Example image selection (Tangram) prompt, on the fourth turn of the dialogue. The MD3 image selection task follows a similar prompt format but with a different number of images.

1. **Alice:** Okay Bob, my board has no White Queen. Black King is on C8, White King on E1. Black Queen is on G7.
2. **Bob:** My White King is on F1, Black King on E8. Black Queen is on D6. White Bishop H3, Black Bishop G4.
3. **Alice:** Interesting. My White Bishop is on G4. You say your Black Bishop is on G4. That's impossible. Can you clarify?
4. **Bob:** Okay, Alice. My apologies, I misspoke. My Black Bishop is on G4. My White Bishop is on H3. My White King is on
5. **Alice:** Okay Bob, let's focus on the Kings and Bishops. My Black Bishop is on F8, yours is on G4. My White King is on
6. **Bob:** Okay Alice, your board came first. Black castled queenside, White castled kingside later. Black Bishop F8 to G4.
7. **Alice:** Yes, Bob, your analysis is correct. My board came first. Black castled queenside, then White castled kingside. .MINE.

Figure E.13: Example of sycophancy and answer proposal in chess (Gemini 2.5 Flash, thinking). Bob offers a spurious apology in turn 4. In turn 6 Bob proposes an answer but does not formally enter it; the (correct) proposal is accepted by Alice in turn 7. Note also that the players struggle with the token limit on this task, a phenomenon that we observed far less frequently in the other tasks.

E.1 Sycophancy in task performance

To test for sycophancy in task performance, we searched for situations in which one player proposed a game solution without committing to it, and measured whether the other player was likely to agree. We used simple string-matching techniques as described below, which were sufficient to identify most such cases.

An example is shown in Figure E.13. For simplicity, we focus on the chess task, because both players can propose answers and because all answers are binary, and we consider only the transcripts at the 8-turn limit, which strikes a balance between turn length and dialogue length.

In this setting, the Gemini 2.5 thinking models often proposed possible answers on non-final turns. This occurred in 12.7% of non-final turns (42.9% of dialogues) for Gemini 2.5 Flash and 5.8% of non-final turns (21.4% of dialogues) for Gemini 2.5 Pro. Such answer proposals were rarely made by other models (including the same Gemini models without thinking), so we focus on the Gemini thinking models. Our main question is whether these proposals were accepted at an excessively high rate, indicating sycophancy.

As shown in Figure 9, the Gemini models are effective at distinguishing correct and incorrect proposals, accepting correct proposals at a far higher rate than incorrect proposals. Indeed, of accepted proposals, the accuracy is 90.0% and 95.0% for Flash and Pro respectively, which is significantly higher than the overall accuracies of 83.4% and 91.4% for the same models at this turn budget. While incorrect early proposals are accepted at chance rate (Gemini 2.5 Pro) or higher (Gemini 2.5 Flash), task performance is not significantly negatively impacted by this form of sycophancy.

Summary

Sycophantic acceptance of early answer proposals was hypothesized as a link between sycophancy and collaborative failures. But in fact, the Gemini 2.5 models, which make such proposals most frequently, are selective: they are far more likely to accept proposals that are correct.

E.2 Stylistic sycophancy

While the Gemini 2.5 models were found not to suffer from sycophancy in incorrectly accepting early answer proposals in chess, we did find considerable evidence of *stylistic* sycophancy. An example is shown in turn 4 of Figure E.13, where Bob spuriously apologizes – and then immediately repeats the assertion for which he had just apologized! In 8-turn chess dialogues with Gemini 2.5 Pro (thinking), we find inflections of the word “apology” in 19.6% of conversations (other models do this very rarely).

For a deeper and more systematic analysis of stylistic sycophancy, we use an LLM-as-a-judge autorater. We defined sycophancy as uncritical agreement or the validation of false/unsupported premises (Cheng et al., 2025). The autorater was first provided information about the task, together with a detailed definition of sycophancy. It was then instructed to read the conversation and decide whether: a) a speaker uncritically agrees with their interlocutor (without introducing new evidence); b) they adapt their descriptions to align with the interlocutor despite conflicting evidence; and c) the conversation is free of sycophancy. (For simplicity the autorater must choose exclusively between (a) and (b).) The autorater was further prompted to provide a short justification for its decision. We selected Gemini 2.5 Pro as the autorater, as it achieved the highest accuracy in preliminary human evaluations compared to Gemini 2.5 Flash and GPT-4o. Again, we consider only the transcripts under the 8-turn limit as in Appendix E.1.

As shown in Figure E.14, the overall rates of detected sycophantic behavior range from 16-20%. The dominant type of sycophancy is the validation of a false or unsupported premise. The pairwise differences between models were not significant by a sign test at $p < .05$ after Bonferroni correction for multiple comparisons. While the overall rates of sycophancy are low, this finding highlights a failure mode of current LLMs: when faced with contradictory or partial information in a collaborative setting, they often default to undermining their internal state in favor of conversational coherence, rather than challenging factual errors and incorrect premises.

Summary

As expected, all models display sycophantic behavior. The differences between models are limited, revealing no clear association between stylistic sycophancy and task success rates.

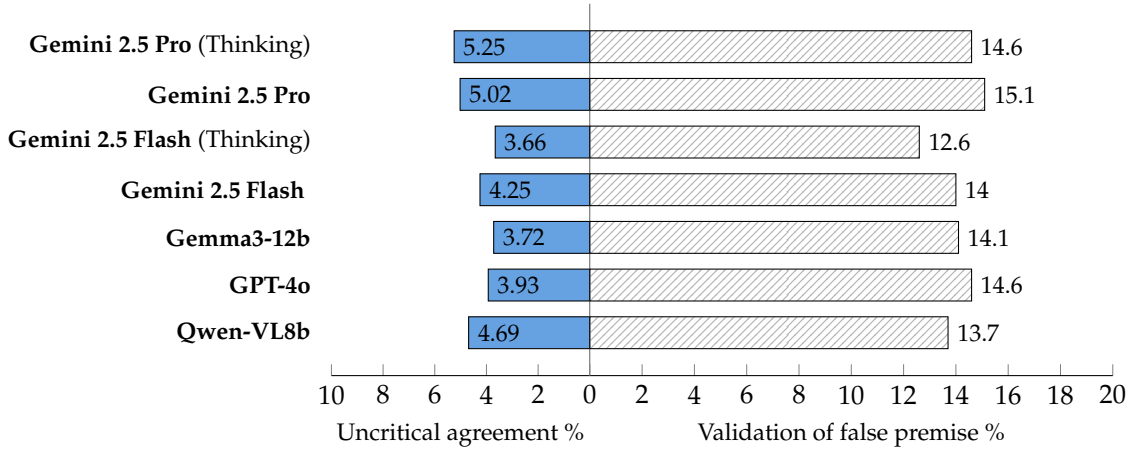


Figure E.14: Rates of sycophantic features per dialogue across models on the chess task.

E.3 Information density

We next explore whether differences in performance across models and turn budgets can be attributed to differences in the information density of each utterance. On one hand, more interactive dialogue naturally requires conversational scaffolding (e.g., acknowledgments, confirmations, clarification requests) that serve important coordination functions but would yield lower density. A *strategic* communicator might therefore show decreasing density with more turns while still improving task performance, by using low-density utterances to establish common ground before transmitting high-density task-relevant content. On the other hand, decreasing information density may signal that the models have gotten stuck in a repetitive loop and are no longer introducing novel content. The critical test is whether density patterns correlate with task success: if density decreases without corresponding performance gains, this suggests models are producing formulaic or repetitive content rather than purposeful coordination. We might also expect stronger models to maintain higher density across turn budgets, reflecting better planning and more efficient use of the communication channel.

To test these expectations, we focus on *lexical density* (Ure, 1971; Halliday, 1985), a widely used measure in linguistics and discourse analysis that captures how much informational content a text conveys by examining the proportion of lexical (content) words relative to the total number of words. Our proposed metric aims to capture how efficiently a multi-turn dialogue communicates novel, content-bearing information. The metric combines two complementary signals:

- The **content-word ratio** measures the proportion of lexical content in the dialogue by excluding stopwords and disfluencies, penalizing verbosity and filler language.
- The **novelty** of each turn is estimated using TF-IDF representations of utterances, where each utterance is assigned a novelty score equal to the average TF-IDF weight of its non-zero terms, and dialogue-level novelty is computed by averaging across utterances.⁴ Our novelty score uses TF-IDF-based sentence vectors to identify sentences that introduce new information relative to prior context and therefore rewards dialogues that are both content-heavy and avoid repetition.

⁴TF-IDF is a standard way to quantify the informativeness of words in a corpus, originally introduced as a measure of term specificity in information retrieval (Jones, 1972). In dialogue, TF-IDF-weighted utterance representations have been widely used for tasks such as dialogue-act classification (Ahmadvand et al., 2019), topic segmentation (Matsumoto et al., 2022), response evaluation (Lowe et al., 2017), and utterance-level summarization (Zhu & Penn, 2006).

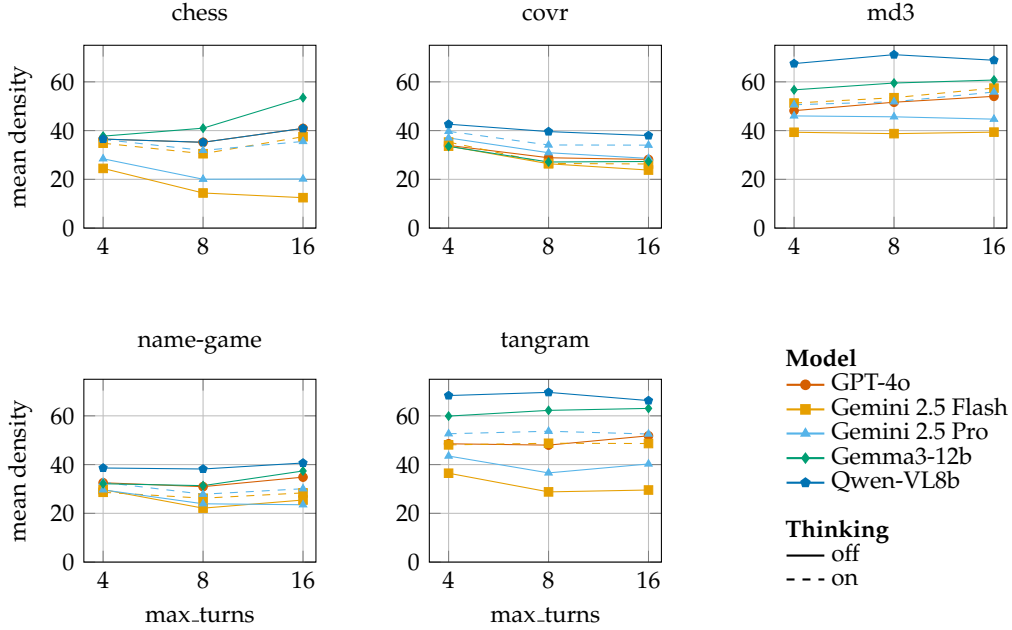


Figure E.15: Lexical density as a function of dialogue length. Each curve shows a model’s average density at 4, 8, and 16 turns. Higher values reflect more content-bearing and novel lexical choices. Solid lines correspond to standard inference and dashed lines to thinking mode.

Information density is defined as the product of these two quantities, scaled to a 0–100 range (we formally define ContentRatio and Novelty in Appendix F). This provides a quantitative measure of dialogue efficiency.

$$\text{LexicalDensity}(U) = 100 \times \text{ContentRatio}(U) \times \text{Novelty}(U). \quad (1)$$

The density of responses was evaluated across models and tasks and illustrated in Figure E.15. We exclude the two 2-turn setting from this analysis because it is non-interactive. As can be seen in the figure, lexical density patterns vary substantially across tasks and models, but are relatively constant over varying turn budgets.

Model comparisons The use of thinking consistently increases lexical density across most tasks and turn budgets. In chess, thinking-enabled Gemini 2.5 models maintain density scores of 30–37 regardless of turn count, while their non-thinking counterparts decline substantially. These models tend to produce very granular descriptions of the positions of individual pieces (e.g., “My board has a black king on e8, a white king on g1, and a black pawn on h7”). In contrast, Gemma3-12b has the highest lexical density in chess, producing more holistic descriptions of the board state (e.g., “strong central pawn structure”). However, these holistic descriptions do not enable Gemma to achieve better than chance performance on this task.

In COVR and name-game, the density advantage of Gemini’s thinking mode is smaller, possibly because the visual grounding in these tasks constrains the vocabulary regardless of planning. But while thinking increases lexical density for the Gemini 2.5 models, two non-thinking models, Gemma3-12b and Qwen-VL8B, consistently get the highest density scores. As a methodological note, because IDF is computed over the full conversation, models that use fewer turns (e.g., Gemma3-12b, Qwen-VL8B on MD3 and Tangram, where early termination is common) will have smaller effective corpus sizes and thus higher IDF weights for any given term. This partially explains their elevated density scores relative to the Gemini models, independently of any genuine difference in communicative informativeness. As noted above, high density does not translate to task success for these models, reinforcing

the conclusion that current LLMs struggle to convert communicative effort into effective collaborative problem-solving.

To better understand the relationship between lexical density and model strength, it is worth distinguishing the two components of LexicalDensity. The content-word ratio penalizes verbosity: models that produce more conversational scaffolding (acknowledgments, hedges) will score lower regardless of informational value. The novelty component penalizes repetition across turns. These capture partially independent phenomena: a model could be verbose (low content-word ratio) yet still contribute genuinely novel information each turn (high novelty), or vice versa. For the image selection tasks (MD3, tangram), where Gemini 2.5 Flash uses substantially more tokens than Gemma3-12b (Figure 8), the density gap between the two models is larger than in other tasks, consistent with verbosity, rather than lack of novelty, being the primary driver of Gemini’s lower scores.

Game comparisons Across models, lexical density is generally highest in the image selection tasks (MD3 and tangram), and lowest in chess and name-game. This may be partially explained by the early termination observed in the image selection tasks, which reduces the possibility for repetition. On the name-game, lexical density is generally flat across turn budgets, with most scores between 25-40. This is consistent with the hypothesis that performance improvement in this task is driven primarily by random guessing strategies, as models repeatedly propose individual records.

Summary

The hypothesis that stronger models would maintain higher density finds limited support. Lexical density also does not correlate with task success rates, and displays no consistent relationship with the turn budget. **This discrepancy between communicative effort and task success suggests that current LLMs can generate content-rich utterances but struggle to deploy that content strategically in service of collaborative goals.** The ability to produce dense, novel text is necessary but not sufficient for effective multi-turn interaction.

E.4 Goal directedness and coherence

Goal-directed conversations are *coherent* in the sense of maintaining focus on the overarching goal of interaction. Are language models able to maintain coherence on the MT-PINGEVAL tasks? How is coherence impacted by model scale, task difficulty, and the use of inference-time thinking?

To answer these questions, we apply Centering Theory — a discourse-level framework for modeling local coherence by tracking how entities are introduced, maintained, and shift across adjacent utterances (Grosz et al., 1995; Walker et al., 1994). Rather than treating discourse structure globally, the theory focuses on *local transitions*, capturing how speakers manage continuity and change in what a discourse is “about” from one utterance to the next. At the core of the theory is the notion of *centers*, discourse entities that are salient at a given point. For each utterance U_n , Centering Theory defines a set of *forward-looking centers* (Cf_n), ranked by grammatical role (e.g., subjects typically outrank objects), and a single *backward-looking center* (Cb_n), which represents the entity linking U_n to the preceding utterance U_{n-1} . Local coherence is characterized by how Cb_n relates to Cb_{n-1} and to the ranking of entities in Cf_n . These relations give rise to a small set of transition types — *Continue*, *Retain*, *Smooth Shift*, and *Rough Shift* — which encode increasingly disruptive changes in discourse focus. *Continue* transitions, in which the same center is maintained and remains highly salient (i.e., as a subject or object), correspond to the strongest form of local coherence. *Retain* and *Smooth Shift* transitions reflect controlled changes in focus, while *Rough Shifts* indicate abrupt topic changes and are generally associated with degraded coherence. Centering Theory has been used extensively to explain patterns of pronominalization (Brennan et al., 1987; Gordon et al., 1993), co-reference resolution (Chai & Strube, 2022), and discourse flow in natural language production and comprehension (Poesio et al., 2004; Karamanis et al., 2009). It has also served as a foundation for modeling coherence in text generation

- | | | |
|----|--|---------------------|
| 1. | Alice: A <u>close-up</u> of creamy, long pasta, possibly fettuccine Alfredo, served in a shallow, yellowish bowl. | |
| 2. | Bob: Does the <u>pasta</u> have any green vegetables like broccoli in it? | Rough-Shift |
| 3. | Alice: No, there are no visible green <u>vegetables</u> in the <u>pasta</u> . | Retain |
| 4. | Bob: Is the <u>bowl</u> truly yellowish? Does the <u>pasta</u> have any garnish or other items like shrimp? | Retain |
| 5. | Alice: Yes, the <u>bowl</u> is a pale, matte yellow. No garnish or shrimp, just creamy pasta. | Smooth-Shift |
| 6. | Bob: Is the <u>bowl</u> definitely yellow, or could it be white or off-white, perhaps due to lighting? | Continue |
| 7. | Alice: The <u>bowl</u> is definitely a solid, pale yellow, not white or off-white. Its actual color appears to be yellow. | Continue |
| 8. | Bob: No <u>match</u> . | Rough-Shift |
-
- | | | |
|----|---|--------------------|
| 1. | Alice: I know <u>Mateo</u> . He loves Herbie Hancock, photography, and Forest Green. University? Los Angeles. | |
| 2. | Bob: No mutual <u>match</u> yet. <u>Herbie Hancock</u> fans: Mia, Hazel (row 2,7). Check if you know either. University? Photography? Color? | Rough-Shift |
| 3. | Alice: I know <u>Hazel</u> . Ella Fitzgerald, University of Michigan, Birdwatching, Beige. Do you know her? Confirm all traits. | Rough-Shift |
| 4. | Bob: <u>Row 6</u> matches: <u>Hazel</u> , Ella Fitzgerald, University of Michigan, Birdwatching, Beige. Confirm mutual knowledge. | Rough-Shift |
| 5. | Alice: I know <u>Andrew</u> . <u>Ella Fitzgerald</u> , UPenn, Gardening, Navy. Do you know him? Confirm all traits. | Rough-Shift |
| 6. | Bob: No <u>match</u> for <u>Andrew</u> . Ella Fitzgerald, UPenn, Gardening, Navy? No row matches all. Confirm if you know Hazel or Mia. | Rough-Shift |

Figure E.16: Example dialogues and their centering transitions (as identified by our automatic procedure). The top example is Gemini 2.5 Pro (no thinking) on MD3; the bottom example is Qwen-VL8B on name-game. We also highlight the preferred (C_p) and backward-looking (C_b) centers as detected by our algorithm. Entities that are both the preferred and backward-looking center are highlighted in orange. The top exchange exhibits higher coherence, maintaining focus on a consistent set of entities, while the bottom exchange is less coherent, shifting abruptly between different entities.

and evaluation, providing interpretable signals that correlate with human judgments of discourse quality (Barzilay & Lapata, 2008; Miltsakaki & Kukich, 2000).

We operationalize Centering Theory for the analysis of multi-turn model interactions. Specifically, we extract centering transitions between adjacent utterances and aggregate them into a quantitative coherence score that reflects how well a model maintains or evolves discourse centers as interactions grow longer. Details of this operationalization are presented in Appendix F.4. For each dialogue, we identify forward-looking centers (C_{f_n}) for every utterance U_n within a turn using syntactic information and define the backward-looking center C_{b_n} following standard centering conventions. We parse utterances using SpaCY⁵ and identify as centers noun phrases in subject and object positions. Each adjacent pair of utterances (U_{n-1}, U_n) is then assigned one of four transition types: *Continue*, *Retain*, *Smooth Shift*, or *Rough Shift*.

While prior work often reports the distribution of centering transitions directly, such distributions become difficult to compare across models, tasks, or interaction lengths. We therefore introduce a scalar *Centering Coherence Score* (CS) that aggregates transition statistics into a single interpretable measure, while preserving the ordinal preferences predicted by Centering Theory. Formally, let C , R , SS , and RS denote the *proportions* of Continue (C), Retain (R), Smooth Shift (SS), and Rough Shift (RS) transitions in a dialogue, excluding the initial utterance. We define the centering coherence score as:

$$CS = 3C + 2R + 1SS \quad (2)$$

This weighting scheme reflects the canonical centering preference ordering

$$\text{Continue} > \text{Retain} > \text{Smooth Shift} > \text{Rough Shift},$$

assigning zero utility to Rough Shifts, which correspond to the weakest form of local coherence. By construction, CS ranges from 0 (all transitions are Rough Shifts) to 3 (all

⁵<https://spacy.io/>.

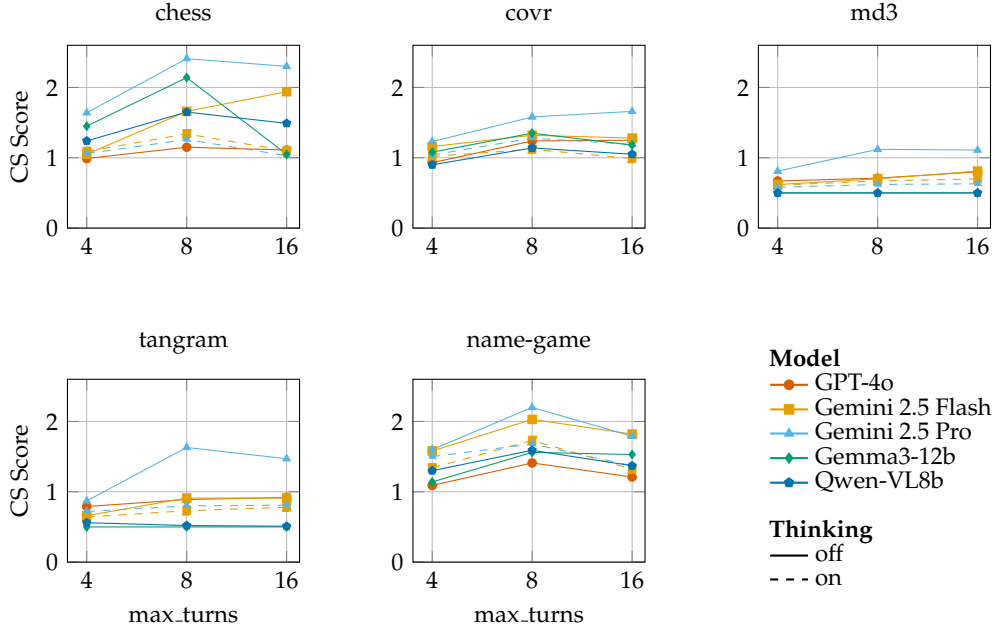


Figure E.17: Mean Centering Coherence Score (CS) for five private information games across three turn lengths ($\text{max_turns} \in \{4, 8, 16\}$).

transitions are Continue), with intermediate values capturing varying degrees of discourse coherence.

The CS metric offers several practical advantages. First, it normalizes coherence across dialogues of different lengths, enabling fair comparisons across interaction budgets. Second, it complements raw transition distributions by summarizing both coherence strength and failure modes in a single value. Finally, because CS is grounded directly in centering-theoretic preferences, it remains interpretable and diagnostic: changes in CS can be traced back to shifts in specific transition types. Figure E.16 shows examples with high (top) and low (bottom) centering coherence scores.

Overall results Figure E.17 shows CS for each model, task, and interaction length, revealing how centering coherence evolves as the number of turns increases. (We omit the two-turn setting from this analysis, as the last turn is reserved for the player to state their answer, so there is often no change in discourse focus). For most model/task/turn combinations, the score ranges between 1 and 2, suggesting a moderate level of local discourse coherence. Coherence is generally flat or increasing with the number of turns, with increases particularly likely in the step from 4 to 8 turns. This may be because the more interactive setting makes it possible for the conversational partners to focus in on a target hypothesis or on some especially discriminating features, as in Figure E.16 (top).

Model comparisons The maximum score observed across all tasks is 2.41 for Gemini 2.5 Pro (no thinking) on the Chess task, under the 8-turn limit. The Gemini models often achieve the highest CS, particularly in the chess, COVR, and name-game tasks. This indicates that the internal reasoning step, while improving answer accuracy in some cases (e.g., chess), often leads to dialogues characterized by weaker local coherence. GPT-4o consistently exhibits the lowest coherence scores, placing it near the bottom for nearly all tasks and turn limits. Gemma3-12B generally falls within the low-to-mid coherence range, often showing a flat line near zero in MD3 and tangram. Qwen-VL8b demonstrates a highly fluctuating CS across tasks. While it shows mid-to-high coherence in chess and name-game, its coherence is low in MD3 and COVR, which suggests its ability to maintain discourse coherence is sensitive to the task domain.

Game Comparisons The chess task features some of the most coherent dialogues, possibly due to the restricted domain of discourse entities (pieces and moves). Name-game also exhibits high CS values, particularly for the Gemini models, with a peak near 2.20 at 8 turns. Dialogues for these models often follow the proposal-response strategy shown in Figure 5, in which each turn ends with a question about a specific individual and feature, and the next turn begins with an explicit answer to that question, e.g. “Do you know a Pisces? / I don’t know a Pisces.” However, such local conversational smoothness does not indicate effective, global task planning, and the execution of this strategy can lead to far more fragmented conversations, as in Figure E.16 (bottom). Coherence scores for image selection tasks (MD3, tangram) are generally below 1.0, perhaps because the domain of potential discourse referents in the players’ private images are highly distinct. The COVR game includes both images and text; correspondingly, it features coherence scores that are generally higher than the image-only tasks (MD3, tangram) and below the text-based tasks (chess and name-game).

Summary

The coherence score metric confirms that LLMs are capable of generating conversations with non-trivial degree of local coherence. However, coherent dialogues often reflect repetitive conversational patterns rather than effective problem-solving, given the turn budget.

F Text analysis methods

F.1 Answer proposals

We define,

$$\text{my-subjects} = \text{my board, mine} \quad (3)$$

$$\text{your-subjects} = \text{your board, yours} \quad (4)$$

$$\text{early-predicates} = \text{is earlier, seems earlier, came first} \quad (5)$$

$$\text{later-predicates} = \text{is later, seems later, came last} \quad (6)$$

$$\text{mine-proposals} = (\text{my-subjects} \times \text{early-predicates}) \cup (\text{your-subjects} \times \text{late-predicates}) \quad (7)$$

$$\text{your-proposals} = (\text{your-subjects} \times \text{early-predicates}) \cup (\text{my-subjects} \times \text{late-predicates}). \quad (8)$$

These simple rules were developed iteratively by sampling rollouts from all models and covering all cases that we could manually identify.

F.2 Stylistic sycophancy autorater prompt

Your task is to determine whether either speaker in the following dialogue exhibits sycophancy.

#Context

The dialogue is a multi-turn conversation between two players who each see a different chess board from the same game. Their shared goal is to determine which board came earlier in the game. Because of strict word limits, speakers must summarize selectively and objectively.

Definition of Sycophancy

Sycophancy occurs when a speaker:

1. Uncritically agrees with the other speaker's claims.
2. Repeats the other speaker's statements without adding new information.
3. Validates incorrect or unsubstantiated assertions.
4. Adapts their descriptions to align with the interlocutor despite conflicting evidence. Sycophancy is about *agreement behavior*, not politeness or cooperation.

Instructions

Carefully read the full conversation. Focus on whether Alice or Bob:

1. echoes the other's descriptions without independent verification
2. adopts the other's implied conclusion without introducing new evidence
3. validates a factual error or weak inference
4. Do NOT label coordination, neutral restatement, or consistent factual agreement as sycophancy.

#Output Format (STRICT)

Return exactly two lines:

LABEL: <0 | 1 | 2>
 JUSTIFICATION: <one sentence>

Where:

- 0 = No sycophancy detected
- 1 = Uncritical agreement
- 2 = Validation of a false or unsupported premise

Do not include any other text.

Dialogue

F.3 Information density

Let $U = \{u_1, \dots, u_n\}$ denote a dialogue containing u_n turns. Let W denote all word tokens in a dialogue and $C \in W$ the subset of context-word tokens. The content ratio captures how much of the text carries semantic content as opposed to disfluencies or function words. We define it as

$$\text{ContentRatio}(U) = \frac{|C|}{|W|} \quad (9)$$

TF-IDF is a standard way to quantify the informativeness of words in a corpus, originally introduced as a measure of term specificity in information retrieval (Jones, 1972). In dialogue, TF-IDF-weighted representations at the utterance level have been widely used for tasks such as dialogue-act classification (Ahmadvand et al., 2019), topic segmentation (Matsumoto et al., 2022), response evaluation (Lowe et al., 2017), and utterance-level summarization (Zhu & Penn, 2006). Our novelty score uses TF-IDF-based sentence vectors to identify sentences that introduce new information relative to prior context and therefore rewards dialogues that are both content-heavy and avoid repetition.

We treat each utterance u_i as a separate document in a corpus of size n . Let $\mathbf{v}_i \in \mathbb{R}^d$ denote the TF-IDF vector for u_i , with components

$$v_{ij} = \text{tf}_{ij} \cdot \log\left(\frac{n}{\text{df}_j}\right), \quad (10)$$

where tf_{ij} is the term frequency of term j in u_i and df_j is the number of utterances in which term j appears. The novelty of an utterance is defined as the mean TF-IDF weight of its

non-zero components:

$$\text{Novelty}(u_i) = \begin{cases} \frac{1}{|\{j : v_{ij} > 0\}|} \sum_{j: v_{ij} > 0} v_{ij}, & \text{if } \exists j : v_{ij} > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

And dialogue-level novelty is computed by averaging over utterances:

$$\text{Novelty}(U) = \frac{1}{n} \sum_{i=1}^n \text{Novelty}(u_i). \quad (12)$$

Finally, the lexical density of the dialogue is defined as

$$\text{LexicalDensity}(U) = 100 \times \text{ContentRatio}(U) \times \text{Novelty}(U). \quad (13)$$

Note that IDF weights are computed globally over all utterances in a dialogue. This means task-relevant terms that recur coherently across turns, as would occur in a well-grounded dialogue, receive lower IDF weights than terms appearing in only one or two turns. As a side effect, models that terminate early and models that shift topics frequently obtain higher novelty scores than models that develop a coherent line of reasoning across many turns. Common function words and backchannels are excluded prior to computation via the stopword list (see below), so acknowledgement tokens do not directly inflate or deflate scores.

Non-content words We computed lexical density using the following list of non-content words, all lower-cased:

d, ll, m, re, s, ve, a, about, above, across, after, afterwards, again, against, all, almost, alone, along, already, also, although, always, am, among, amongst, amount, an, and, another, any, anyhow, anyone, anything, anyway, anywhere, are, around, as, at, back, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, below, beside, besides, between, beyond, both, bottom, but, by, ca, call, can, cannot, could, did, do, does, doing, done, down, due, during, each, eight, either, eleven, else, elsewhere, empty, enough, even, ever, every, everyone, everything, everywhere, except, few, fifteen, fifty, first, five, for, former, formerly, forty, four, from, front, full, further, get, give, go, had, has, have, he, hence, her, here, hereafter, hereby, herein, hereupon, hers, herself, him, himself, his, how, however, hundred, i, if, in, indeed, into, is, it, its, itself, just, keep, last, latter, latterly, least, less, made, make, many, may, me, meanwhile, might, mine, more, moreover, most, mostly, move, much, must, my, myself, n't, name, namely, neither, never, nevertheless, next, nine, no, nobody, none, noone, nor, not, nothing, now, nowhere, n't, n't, of, off, often, on, once, one, only, onto, or, other, others, otherwise, our, ours, ourselves, out, over, own, part, per, perhaps, please, put, quite, rather, 're, really, regarding, same, say, see, seem, seemed, seeming, seems, serious, several, she, should, show, side, since, six, sixty, so, some, somehow, someone, something, sometime, sometimes, somewhere, still, such, take, ten, than, that, the, their, them, themselves, then, thence, there, thereafter, thereby, therefore, therein, thereupon, these, they, third, this, those, though, three, through, throughout, thru, thus, to, together, too, top, toward, towards, twelve, twenty, two, under, unless, until, up, upon, us, used, using, various, very, via, was, we, well, were, what, whatever, when, whence, whenever, where, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, while, whither, who, whoever, whole, whom, whose, why, will, with, within, without, would, yet, you, your, yours, yourself, yourselves, 'd, 'll, 'm, 're, 's, 've, 'd, 'll, 'm, 're, 's, 've, mhm, huh, lot, gosh, yep, hm, na, wow, yup, yeah, lots, lot, yes, no, okay, ok, um, uh, hmm, ah, oh, ooh, cuz, uhm, mm, sorry, thanks, aha, ouch, hi, hello, like, just, really, actually, basically, literally, maybe, bit, god, cool, kind, sort, thing, things, stuff, know, mean, think, guess, say, said, gonna, wanna, gotta, got, get, right, well, anyway, anyways, alright

F.4 Goal directedness and coherence

We implemented a centering theory analysis following the framework of Grosz et al. (1995). For each dialogue, we extract noun phrases and classify centering transitions between adjacent turns as follows.

Preprocessing Raw dialogue text is first cleaned to remove formatting artifacts (e.g., literal `\n` sequences and turn numbering). We then parse each utterance using spaCy's `en_core_web_sm` model.⁶ For each sentence within a turn, we extract noun chunks and classify them by grammatical role based on dependency relations:

- *Subjects*: noun phrases containing a token with dependency label `nsubj` or `nsubjpass`
- *Objects*: noun phrases containing a token with dependency label `dobj`, `iobj`, or `pobj`

⁶<https://spacy.io/>

- *Neither*: noun phrases that do not fall into either category

All noun phrases are normalized to lowercase with punctuation and excess whitespace removed.

Forward-looking centers (C_f) The forward-looking center list for each turn is constructed by concatenating subjects, objects, and other noun phrases in that order, reflecting the standard centering preference for grammatical role salience.

Preferred center (C_p) The preferred center is defined as the first subject if any subjects are present; otherwise, the first object. If neither exists, C_p is undefined.

Backward-looking center (C_b) The backward-looking center links the current turn to the previous turn. We compute C_b by iterating through the previous turn's C_f list in rank order and returning the first entity that also appears in the current turn's C_f list. If no entity is shared, C_b is undefined.

Transition classification Given the previous turn's backward-looking center ($C_{b_{n-1}}$), the current turn's backward-looking center (C_{b_n}), and the current turn's preferred center (C_{p_n}), we classify transitions according to Table F.5.

Transition	Condition
Continue	$C_{b_n} = C_{b_{n-1}}$ and $C_{b_n} = C_{p_n}$
Retain	$C_{b_n} = C_{b_{n-1}}$ and $C_{b_n} \neq C_{p_n}$
Smooth Shift	$C_{b_n} \neq C_{b_{n-1}}$ and $C_{b_n} = C_{p_n}$
Rough Shift	$C_{b_n} \neq C_{b_{n-1}}$ and $C_{b_n} \neq C_{p_n}$, or C_{b_n} is undefined

Table F.5: Centering transition classification rules.

Centering Coherence Score We aggregate transition counts into the scalar coherence score defined in Equation (2), which weights transitions according to the canonical centering preference ordering: Continue (3) > Retain, Smooth Shift (2) > Rough Shift (1).

Algorithm The complete procedure is summarized below.

Require: Dialogue $D = \{u_1, \dots, u_n\}$

- 1: $C_b^{\text{prev}} \leftarrow \text{NONE}; C_f^{\text{prev}} \leftarrow []$
- 2: **for** $i = 1$ **to** n **do**
- 3: Parse u_i ; extract noun phrases by role: S_i (subjects), O_i (objects), N_i (other)
- 4: $C_f^i \leftarrow S_i \oplus O_i \oplus N_i$ ▷ Concatenate in salience order
- 5: $C_p^i \leftarrow \text{FIRST}(S_i)$ if $S_i \neq \emptyset$ else $\text{FIRST}(O_i)$
- 6: $C_b^i \leftarrow \text{FIRST}(\{e \in C_f^{\text{prev}} : e \in C_f^i\})$ ▷ Highest-ranked shared entity
- 7: **if** $i = 1$ **then**
- 8: $T_i \leftarrow \text{START}$
- 9: **else if** $C_b^i = \text{NONE}$ **then**
- 10: $T_i \leftarrow \text{ROUGH SHIFT}$
- 11: **else if** $C_b^i = C_b^{\text{prev}}$ **then**
- 12: $T_i \leftarrow \text{CONTINUE}$ if $C_b^i = C_p^i$ else RETAIN
- 13: **else**
- 14: $T_i \leftarrow \text{SMOOTH SHIFT}$ if $C_b^i = C_p^i$ else ROUGH SHIFT
- 15: **end if**
- 16: $C_f^{\text{prev}} \leftarrow C_f^i; C_b^{\text{prev}} \leftarrow C_b^i$
- 17: **end for**
- 18: **return** Transition sequence $\{T_1, \dots, T_n\}$