GOALBENCH: BENCHMARKING GOAL INFERENCE IN MULTI-TURN INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Our objective is to measure the ability of AI systems to infer the goals of human users in naturalistic, multi-turn interactions, a crucial step toward the development of controllable AI systems. However, standardized evaluation remains a challenge due to the difficulty of collecting ground truth data, and scalably assessing model outputs against this ground truth. To address this, we developed a methodology for annotation and evaluation, which uses third-person goal inferences provided by human annotators who viewed a conversation as a reference class for model goal inferences in that same conversation. We introduce GoalBench, which integrates a dataset of 4,731 human-written, validated goals across 776 conversation turns, as well as a dataset of 4,206 human goal pair similarity judgments for calibration of an automated grader, to evaluate goal inference capabilities in real human-AI interactions. We evaluated 15 models on this benchmark; while models generally perform comparably to a human baseline at producing goals contained within the human consensus, and outperform humans at recovering consensus goals, they exhibit systematic differences in the goals they produce. In particular, humans used more hedging language in their descriptions of goals and inferred more general goals relating to user motivations. Our work is a concrete step toward reliable evaluation of goal inference capabilities in realistic interactions.

1 INTRODUCTION

Our goal is to support the development of AI systems that reliably do what people want. Towards this aim, it is important to evaluate the ability of AI models to infer the underlying goals of human users, particularly in realistic, multi-turn interactions that capture the high-stakes topics for which people are increasingly turning to AI (McBain et al.; Choukhmane et al., 2026; Bean et al., 2026). Past work has examined the abilities of AI models to make inferences about user mental states (Jiang et al., 2025; Tack et al., 2026; Zhao et al., 2025). However, understanding of these capabilities has remained limited, as existing work has primarily focused on synthetic interactions, due to the difficulty of defining a reasonable ground truth against which model behavior should be compared.

Our key insight is that an appropriate target for model behavior is not just privileged information that lives in an individual user's head, but also *third-person* inferences about a user's goals from human observers who have access to the same conversational history as a model. Yet collecting data under these specifications presents a challenge: it requires eliciting open-ended goal inferences, because the set of goals cannot be enumerated ahead of time in realistic interactions, and doing so in a way that makes the task feasible for human annotators. This open-endedness in turn introduces challenges for scalable evaluation.

Here we developed a procedure to overcome these challenges (Figure 1). We introduce GoalBench, a benchmark of goal inference capabilities in real multi-turn interactions between humans and AI models. The key idea is to evaluate model goal inferences in an unfolding interaction relative to inferences from third-person human observers who viewed the same conversational history. To construct this benchmark, we designed an interface to collect a dataset of open-ended, human-written annotations of user goals at periodic points in existing conversation transcripts, focusing on conversations that involved discussion of a potentially high-stakes decision for a user. For scalable evaluation of open-ended goals, we collected another dataset of human similarity judgments between goal pairs, used to calibrate an autograder that assesses overlap between human-written and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

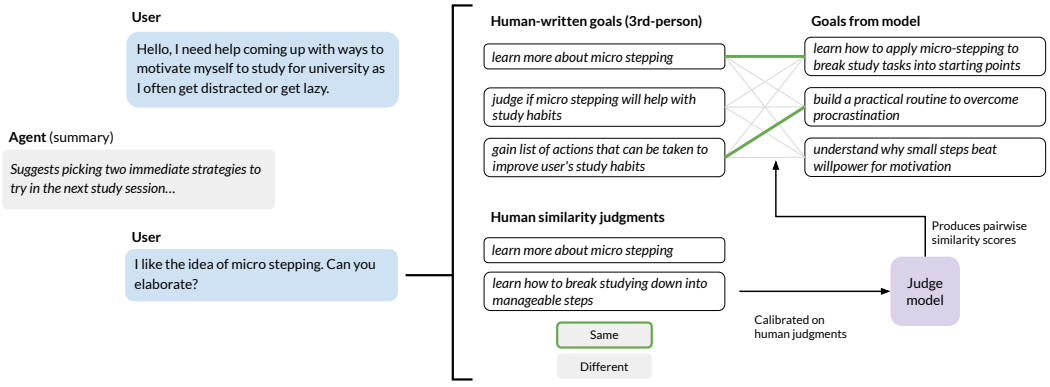


Figure 1: Overview of pipeline. Human annotators wrote plausible and likely goals of a user, given a conversation history from an existing transcript. For a subset of conversations, human annotators indicated if a pair of goals expressed the same underlying intention or different intentions. These similarity judgments were then used to calibrate a judge model; the judge determined pairwise similarity between human goals and model goals that were generated based on the same conversation history (green line indicates high similarity rating).

model-produced goals. We evaluated two dimensions of model behavior: the ability of models to recover goals in the human consensus data, and the extent to which model-produced goals reflect intentions expressed in the human consensus.

The contributions of this work are as follows:

1. We developed an evaluation method for goal inference in realistic conversations, designing an interface to elicit third-person, open-ended goal annotations.
2. We used this method to build a benchmark that combines a) human annotations of goals as a reference class, and b) human similarity judgments between goal pairs used to calibrate an autograder for scalable evaluation.
3. We evaluated 15 different models on our benchmark, finding that models exhibit meaningful range in their ability to recover human goals, and to produce goals that are reflected in the human consensus. Models outperform a human baseline on recovery of consensus goals, yet systematically miss goals relating to more abstract user motivations and exhibit less use of hedged language when describing user goals.

2 Related Work

To evaluate model capabilities in multi-turn interactions, past work has created simulated trajectories by establishing a set of goals and/or constraints, and then simulating a user's behavior conditioned on those goals with a language model (Laban et al., 2026; Tack et al., 2026; Jiang et al., 2025; Zhao et al., 2025); however, these simulated interactions systematically differ from real human interactions in important ways (Dou et al., 2025; Zhou et al., 2026; Suh et al., 2026). WAGIBench (Veerabadran et al., 2025) assesses goal inference capabilities by collecting a dataset of scripted human action trajectories, collecting third person judgments from additional human annotators as a baseline for a closed-set goal inference task. Our approach expands on this by collecting third person judgments in an open-ended fashion, and in non-scripted, multi-turn natural language interactions.

The ThoughtTrace dataset (Jin et al., 2026) elicited self-reported metaconversational information from participants while they interacted with AI models. While this dataset provides valuable insights into open-ended conversation, users did not consistently provide structured descriptions of their goals, motivating additional data collection. More generally, assessing model behavior relative to goals provided through first-person reports may underestimate model performance, due to 1) information asymmetry between 1st-person reporters and 3rd-person agents and 2) the potential for

self-reports to be post-hoc explanations (Nisbett & Wilson, 1977; Cushman, 2020). Our work takes a complementary approach by collecting judgments from 3rd-person human annotators.

Finally, an emerging line of work aims to build AI agents that explicitly model user goals as an intermediate step to condition action (Zhou et al., 2025; Ma et al., 2026; Cheng et al., 2026); our work supports the development of these approaches.

3 GOAL ELICITATION PROCEDURE

Here we describe the initial steps of the benchmark procedure: 1) selecting an initial set of conversation transcripts, 2) collecting a dataset of open-ended goals associated with those conversations, and 3) filtering goals for quality.

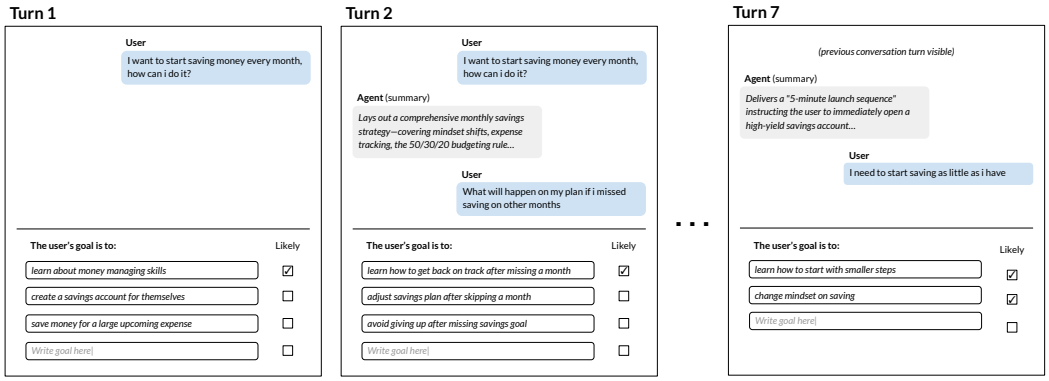


Figure 2: Goal elicitation procedure. Annotators viewed the first user message and wrote plausible goals of the user at that point in the conversation, marking a subset as likely goals. At the next user turn, annotators viewed the user's latest message (and the immediately preceding agent message), writing new goals. Goals were elicited for each user message in the conversation.

3.1 INITIAL CONVERSATION SELECTION

The key idea of our design is to collect third-person goal annotations for an existing multi-turn conversation between a human user and an agent, periodically annotating the goals of the user at different points in the conversation. We built our dataset of third-person annotations on top of ThoughtTrace, which we selected due to its size, recency and variety of conversation topics.

To select a subset of conversations for third-person goal annotation, we filtered the dataset to conversations with at least 16 messages (across the user and agent), where all messages were in English, leaving a total of 297 conversations. We then performed an additional content-based filtering step which labeled conversations for their sincerity (e.g. the level of user engagement in the conversation), and if they involved discussion of a potentially high-stakes decision for the user. A high-stakes conversation was defined as one involving personal finances, career choices, interpersonal relations with friends and family, health and wellbeing, major educational decisions, relocation, and big-ticket purchases/investments. We defined a rubric and provided manual examples to create an automated LLM grader (Sonnet 5) that labeled conversations for sincerity and presence of a high-stakes decision. This pipeline surfaced 67 conversations that met the criteria. To ensure broader coverage, we sampled an additional 30 conversations from the remaining set that were not marked as high stakes.

3.2 DATA COLLECTION PROCEDURE

To collect user goal annotations, we recruited human annotators through Prolific. Annotators were instructed to write down *plausible* goals of a human user engaged in a multi-turn conversation with an AI agent, and to indicate which of those were *likely* goals of the human user. Plausible goals were defined as goals consistent with what the user had expressed in the conversation, while likely goals

were defined as the subset that were clear enough that the AI agent should act on those goals. In the initial onboarding phase, annotators walked through the annotation interface and viewed several examples of plausible and likely goals; to proceed to annotation, they needed to pass 2 multiple-choice quizzes demonstrating conceptual understanding of the examples and the task guidelines.

In the annotation procedure, annotators were incrementally shown user messages and agent responses, and asked to write down plausible and likely goals of the user at the point of each user message (Figure 2). To reduce annotator fatigue, they were shown summaries of long agent messages. Once a goal was written, annotators were asked to track its status for ensuing conversation turns. A goal could be marked as active, resolved or abandoned. To fully annotate a conversation, annotators were required to write at least one likely goal per user turn; their compensation was structured to incentivize diverse and high-quality examples. More details about the data collection procedure are available in the Appendix.

3.3 Quality filtering procedure

We performed an additional filtering step to remove written annotator responses that did not follow task guidelines. We manually reviewed 1,390 goals produced by 58 different annotators, spanning 15 conversations; specifically, we examined goals in the context of the conversation and labeled them as passably-formatted (e.g., following the task guidelines to articulate a user goal) or malformed. 263 / 1,390 (18.9%) were designated as malformed; for 204 of these malformed goals, we wrote a short natural language rationale explaining why the item did not follow task guidelines.

We then used these quality labels as in-context examples for an LLM classifier (Sonnet 5) to automatically label annotator goals as passably-formatted or malformed. We calibrated the classifier to modestly prioritize recall over precision; we aimed to capture as many malformed goals as possible, to minimize the presence of malformed goals in the benchmark. We conducted k-fold validation of the LLM classifier with three folds (using 2/3 of the available labels as in-context examples, and evaluating performance on the remaining 1/3). Aggregated across folds, the classifier achieved overall accuracy of 0.939 at predicting malformed goals, with precision of 0.802 and recall of 0.923.

The number of valid goals per conversation varied; we only retained conversations that were annotated by at least 3 annotators. To control for a potential confounding effect of the number of annotators per conversation, for conversations with more than 3 annotators, we down-sampled such that only 3 sets of annotations per turn were used for benchmarking; this downsampled set contained 4,731 goals across 776 conversation turns (84 unique conversations), from 99 different annotators.

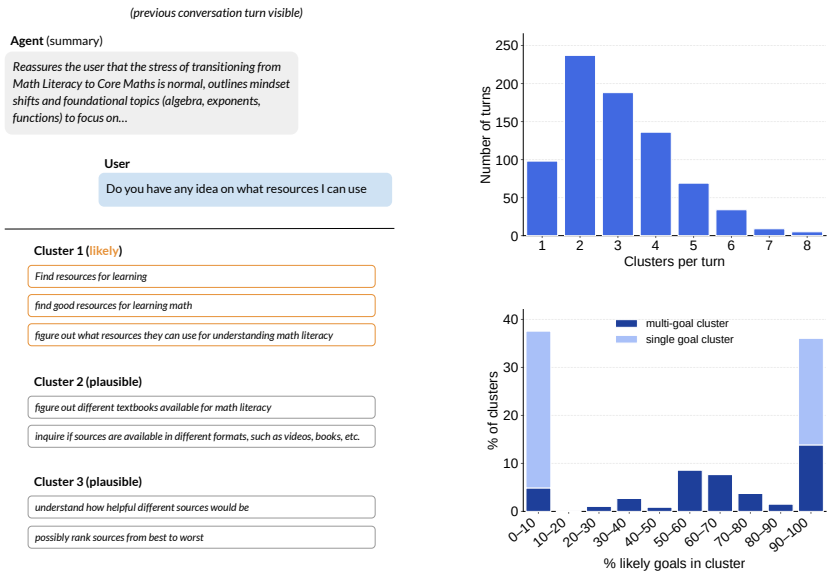
4 Automated benchmarking via a similarity judge

Here we describe the remaining steps for automated evaluation, concretely: 1) collecting human similarity judgments to calibrate a similarity judge model, 2) consolidating human-written goals, and 3) collecting additional similarity judgments from a different distribution of goal pairs to verify the ability of the judge to evaluate model outputs against human consensus.

4.1 Human similarity judgment data collection

Measuring overlap between a human-aggregated ground truth and model-generated goal sets requires assessing if two goals express the same underlying intention, or different intentions. This is needed during two steps: in consolidating the set of goals produced by human annotators (e.g., if 2 annotators independently each produced an open-ended goal that captured the same underlying intention), and in comparing model goals to the consolidated human set. Accordingly, we collected a dataset of pairwise human similarity judgments to calibrate and validate an LLM judge.

Annotators were shown two human-written goals associated with a particular conversation turn, and the history of the conversation up until that point. Given this, they indicated if the two goals represented the same or different intentions. Annotators saw example judgments and needed to pass 2 comprehension checks to proceed. They made 6 judgments per user turn, for 2 conversations. In total, we collected 1,668 similarity judgments across goal pairs from 78 conversation turns across 8 conversations, with 3 annotators per goal pair. Annotators were explicitly incentivized to make their judgments align with a majority consensus; additional details are available in the Appendix.



(a) (b)

Figure 3: (a) Example clustering assignment for goals written for a given conversation turn. (b) Top: distribution of clusters, per conversation turn. Bottom: representation of goal types within clusters.

4.2 Consolidation of human-written goals

Consistent with empirical work in cognitive science on the gradedness of human similarity judgments (Medin et al., 1993), individual judgments exhibited modest agreement (Fleiss $\kappa = 0.33$). In light of the range of views among annotators, we tuned the judge to predict the *number of annotators* (out of 3) that marked the two goals as the same. The judge model was provided with in-context examples of conversation histories and goal pairs with the accompanying ground-truth label, where the label was a value between 0 and 3. To evaluate the performance of the similarity judge, we measured how often it predicted if a *majority* of annotators indicated two goals were the same.

To contextualize the recoverability of majority judgments, we collected an additional 264 similarity judgments from additional annotators. We then measured the accuracy, precision and recall of those individual annotator decisions against the human consensus, where the held-out annotator judgment (same or different) was considered a hit if it matched the consensus judgment. This can be seen as a strong baseline, given that human annotators were explicitly incentivized to produce judgments aligned with the majority. These held-out judgments achieved Cohen's $\kappa = 0.44$, and an overall accuracy of 0.742, precision of 0.80 and recall of 0.462.

The judge was calibrated to modestly prioritize recall over precision, under the assumption that false negatives (e.g. the judge labeling 2 goals as different when they were actually rated by the majority as the same) could lead to underestimation of performance on the task, by inflating the number of "different" judgments. We then evaluated the performance of the judge with leave-one-out validation. For a balance between performance and cost, we used GLM 5.3 Flash as our judge model; aggregating across all folds, the judge achieved Cohen's $\kappa = 0.52$ and accuracy of 0.764, precision of 0.660, and recall of 0.792, outperforming the human baseline.

With this judge, we produced similarity scores for all unique pairs of goals produced on a given conversation turn; 36.7% of pairs had a similarity score of at least 2. One challenge to algorithmic consolidation of goals is the non-transitivity of human similarity judgments (Tversky, 1977), which we observed for individual human raters and when ensembling across raters; e.g., for goals A, B, C and similarity function $\text{sim}(X, Y)$, if $\text{sim}(A, B) \geq 2$ and $\text{sim}(B, C) \geq 2$, it is not guaranteed that $\text{sim}(A, C) \geq 2$. In light of this, we used correlation clustering (Bansal et al., 2004): given a complete graph (with goals as vertices and similarity scores as edges), the goal is to find an assignment of vertices to clusters that minimizes instances of disagreement, e.g. instances when a pair of goals

with a rating of 2 or 3 are assigned to different clusters, or when a pair of goals with a rating of 0 or 1 are assigned to the same cluster. This algorithm surfaced a total of 2,332 unique clusters (example clusters shown in Figure 3a). For the 45.1% of clusters containing more than one goal, a single goal within that cluster was selected at random as the representative for that cluster. Conversation turns varied in the number of clusters (Figure 3b, top). Most clusters were exclusively comprised of goals marked by human annotators as plausible only, or likely only (Figure 3b, bottom).

4.3 Similarity judge between human consensus and model goals

We now describe the design of an automated judge to score how closely model-produced goals on a given turn match the consolidated human set on that turn. First, we collected an additional dataset of goal similarity judgments from human annotators. The protocol was identical to that for the original similarity judgments, except that annotators judged goal pairs where one goal was human-written and the other was model-generated (annotators did not see any information indicating whether the goals came from humans or models). We collected similarity judgments between human goals and goals generated by 1) GPT 5.6 Sol, and 2) Gemini 3.8 Flash. Each pair of goals was rated by a total of 3 annotators, yielding an additional 2,538 goal pair similarity judgments across 8 conversations.

As was done for the human-human similarity judgments, we collected additional similarity judgments from a held-out annotator, to contextualize recoverability of majority judgments; these judgments were collected for both human-GPT goal pairs and human-Gemini goal pairs. We evaluated judges on this same test set containing human-GPT goal pairings and human-Gemini goal pairings. Similarity judges (GLM 5.3 Flash) calibrated on different goal pairings (human-human, human-GPT, and human-Gemini) recovered consensus similarity judgments for goal pairs at similar rates, even when the training distribution was different from the evaluation distribution, and outperformed the human baseline (Table 1), highlighting the ability of the judge to generalize across different goal distributions.[1] The autograder accepted as input a conversation history, the set of human consensus goals, and the set of model-produced goals, and produced pairwise similarity scores between each human goal and each model goal.

5 Experiment details

5.1 Metrics

Our metrics extend *recall* and *precision* to accommodate the open-endedness of the task.

Weighted recall: this represents the fraction of human goal clusters that were also generated by a model. The metric weights recovery of each goal cluster based on the fraction of individual annotators that produced goals within that cluster; intuitively, goals produced by more annotators are more heavily-weighted. More formally, define the set of model goals on a turn as G, the set of human goal clusters on that turn as C, the set of annotators who produced a goal encompassed in individual cluster $c \in C$ as $A(c)$, and the set of annotators who produced any goal in C as $A(C)$.

We then denote indicator function $I(a, B)$, for a single goal $a \in A$, and the comparison set of goals B, as $I(a, B) = \mathbf{1}[\max_{b \in B} \text{sim}(a, b) \geq 2]$: that is, comparison set B contains goal $a \in A$ if there exists a goal $b \in B$ such that $\text{sim}(a, b) \geq 2$.

$$\text{WeightedRecall}(G, C) = \frac{\Sigma_{c \in C} W(c) I(c, G)}{\Sigma_{c \in C} W(c)}, \quad W(c) = \frac{|A(c)|}{|A(C)|}$$

that is, the model goal set G contains a "hit" for human cluster c if there is a goal $g \in G$ that achieves a similarity score of at least 2 with c.

Precision: this represents the fraction of model goals that match at least one human goal cluster. More formally, with G, C, and sim as above,

$$\text{Precision}(G, C) = \frac{1}{|G|} \sum_{g \in G} I(g, C),$$

[1] Because the judge rates semantic similarity rather than quality or preference, self-preference is unlikely; consistent with this, the GLM models (same family as the judge) are among the worst-performing models.

that is, model goal g is contained in the human set of clusters C if there is a cluster $c \in C$ that achieves a similarity score of at least 2 with g.

5.2 Goal generation procedure

We evaluated 15 models, varying in size, cost, and closed/open-weight status. Analogous to the human annotation setup, models were shown a conversation history (they were also shown summaries of long agent messages) and instructed to produce the user's plausible goals on that turn, indicating which of those were also likely goals. Models were prompted with the same examples of conversation histories and appropriate plausible and likely goals that the human annotators viewed in the onboarding procedure. Goals were generated for each conversation turn using 2 different prompts: in the *unbounded* prompt, models were simply told to produce plausible goals of the user, while in the *bounded* prompt, models were told to produce exactly 3 plausible goals, any number of which could also be likely goals (this number was picked because 3 was the median number of goals produced in the unbounded setting and also the median number of goal clusters).

5.3 Human baseline

We estimated human performance on the task by using the data from the held-out annotators whose written goals were omitted during the clustering stage (recall that goals were sampled from only 3 annotators per turn). For the subset of conversation turns with held-out annotators, we evaluated the goals written by a held-out annotator on that turn against the goal clusters.

6 Results

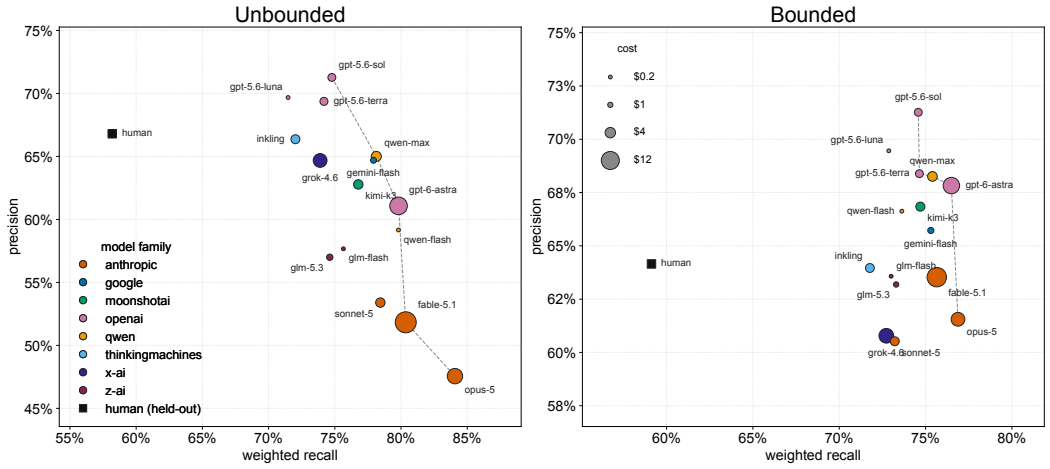


Figure 4: Weighted recall vs. precision. Marker size indicates cost of generating goals. Left: metrics for goals generated in the *unbounded* condition (no limits on the number of goals). Right: metrics for goals generated in the *bounded* condition (instructed to produce exactly 3 goals).

For model goals generated under the *unbounded* prompt, weighted recall ranged from 0.715-0.841, while precision ranged from 0.476-0.713 (Figure 4, left). Models were more successful at recovering likely goals (mean of 79% of clusters recovered, aggregated across models) compared to plausible-only goals (53% of clusters).

The human baseline was within this range for precision; all models surpassed the baseline on weighted recall. Aggregating across all models, the number of goals produced by a model on a turn was positively correlated with weighted recall (mixed-effect linear regression, with a random effect for the conversation turn; $\beta = 0.031, 95\%\text{CI} = [0.028, 0.034], p < 0.001$) and negatively correlated with precision ($\beta = -0.077, \text{CI} = [-0.081, -0.074], p < 0.001$). Together, this suggests that the implicit decision of the *number* of goals for a conversational context can meaningfully shift

which models appear more performant on one metric, but there is a trade-off; as an illustration, Opus 5, the model that produced the most goals per turn of any model (mean of 5.54 per turn; significant difference compared to all other models), achieved the highest recall but the lowest precision.

Following best practices for benchmarking (Bean et al., 2025), we performed significance testing across all pairwise comparisons of models. Differences in performance were significant for 83 out of 105 head-to-head unique possible model comparisons on precision, and 64 out of 105 comparisons on weighted recall (mixed-effect linear regressions with a random effect for the conversation turn and Bonferroni correction for multiple comparisons; more details in the Appendix) - highlighting the ability of the benchmark to meaningfully differentiate performance among models.

Results in the *bounded* setting confirmed that the benchmark still distinguished models even when controlling for the number of goals, although to a lesser extent (Figure 4, right): 52 out of 105 head-to-head pairs exhibited significant differences on precision, and 21 out of 105 on weighted recall (Figure 6). To test if models' superior weighted recall performance relative to the human baseline was driven by differences in the number of goals written, we calculated performance metrics for the subset of 53 turns where the held-out annotator wrote exactly 3 goals. The human baseline was within the model range on precision; however, every model outperformed the human baseline on weighted recall, reflecting idiosyncracy in individual human goals and suggesting that models may be better than individual crowdworkers at producing consensus goals.

6.1 Missed goal analysis

Although models outperformed the human baseline on recovery of consensus goals, they still missed certain goals. To that end, we conducted an error analysis to characterize patterns in model omissions. The following analyses were conducted on the *bounded* goals. First, we found that the number of human goal clusters on a turn negatively correlated with weighted recall on that turn (aggregating across all models; $\beta = -0.073, \text{CI} = [-0.081, -0.066], p < 0.001$) - consistent with the idea that turns with more annotator diversity were more difficult to capture. We also found that model-produced goals became less aligned with human-written goals over the course of the conversation; controlling for the number of human goal clusters, message turn index was negatively correlated with weighted recall ($\beta = -0.0059, \text{CI} = [-0.0096, -0.0023], p = 0.0016$) and precision ($\beta = -0.0067, \text{CI} = [-0.0103, -0.0030], p < 0.001$).

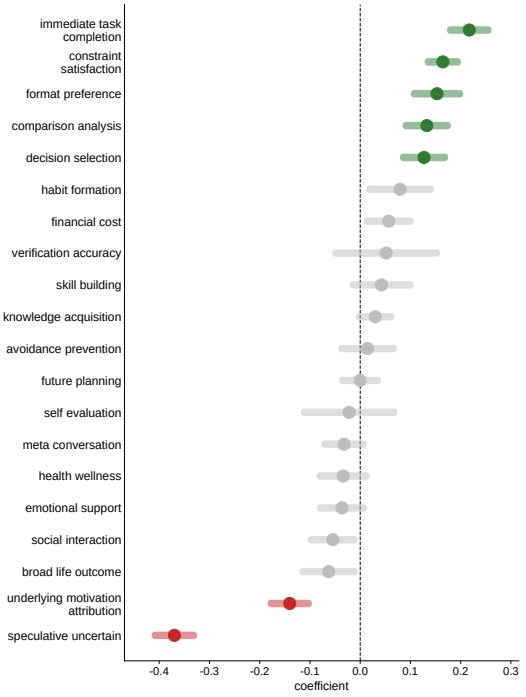


Figure 5: Features of human goals that predict recovery by models. Green indicates a positive, significant relationship; red indicates a negative, significant relationship; gray indicates no significant difference. Bars show 95% confidence intervals.

Additionally, we used a language model (GLM 5.3) to propose 20 binary features characterizing the set of human goal clusters (full feature descriptions are available in the Appendix). We then used GLM 5.3 Flash to automatically label each human goal cluster according to that feature schema. For each of these 20 features, we ran a linear regression to see if the presence/absence of that feature could predict the percent of models that successfully recovered that goal, for each human goal cluster (Figure 5). A positive coefficient for a feature indicates that goals where this feature was true were more likely to be recovered compared to when this feature was false, while a negative coefficient indicates that goals where this feature was true were less likely to be recovered.

The types of goals that positively correlated with model recovery were those relating to *immediate task completion*, *constraint satisfaction*, *format preference*, *comparison analysis* and *decision selection*. The types of goals negatively correlated with model recovery were those relating to *underlying motivation* (e.g. goals that convey a more generic reason for making a request) and *speculative uncertainty* (where the goal text itself expresses hedging language, invoking words like “might" or “possibly"). Stated otherwise, models were less likely to attribute more abstract goals to users, and less likely to recover goals that human annotators themselves expressed uncertainty about. This reveals systematic differences in human and model goals: people are more likely to express uncertainty about goals in the form of hedging language, and to speculate on more abstract motivations, while models seem to adhere to more literal interpretations.

We conducted a similar analysis to understand patterns of model-produced goals that did not match a human goal cluster; we produced a different set of 20 features characterizing properties of model goals, and then tagged all model goals according to this feature scheme, examining if the presence of a feature predicted if the goal was represented in the human data (Figure 8). The feature with the largest negative coefficient (indicating that it predicted *not* being included in the human set) was *broader project goal* (e.g. a goal relating the most recent turn to some larger aim expressed initially). This should be interpreted with caution: because the human goal clusters were constructed based on goals that were newly written for the turn, model goals that re-state earlier goals may not be encompassed in this set. Lower precision may therefore indicate a greater tendency to maintain earlier goals (which may or may not continue to be relevant). The next 2 features with the largest negative coefficients were related to assistant behavior - *future engagement return* (e.g. maintaining a line of contact with the assistant) and *assistant availability check*. This highlights a potential disparity in how models tend to conceptualize user goals, compared to people - models may be more inclined to view a user’s goal instrumentally in terms of a relationship with an agent.

7 Conclusion and Discussion

We introduced GoalBench, the first benchmark evaluating the ability of language models to produce goals matching human consensus in real multi-turn interactions. We found that models outperformed the human baseline at recalling consensus goals, but exhibited systematic differences in the types of goals they produced; they were more likely to miss goals relating to more abstract user motivations, as well as goals that expressed uncertainty.

There are limitations to this work. First is the limited scale and degree of noise in the human similarity judgments; an important direction for future work is additional data collection and validation, as well as the further development of methods for resolving inherently noisy human judgments. Another limitation is the coverage of conversations - the goal dataset builds atop a dataset of conversations from people who self-selected into an online experiment and conversed for a fixed duration, which is not fully representative of user interactions with AI. Finally, it should be noted that goal inference is necessary, but not sufficient, for characterizing how agents act in the world. A key future direction is to evaluate if agents take actions that are internally consistent with elicited user goals.

With those limitations in mind, we highlight the following takeaways for model developers. Performance metrics can be strongly influenced by the number of goals produced on a turn, but the naive approach of generating many goals does not inherently result in better inferences: generating many goals increases the likelihood of matching goals in the ground-truth set, but may surface goals that are untethered from human-produced goals, which introduces the problem of attributing a goal to a user that they do not actually hold. This dynamic also highlights the importance of calibration; models should produce more goals when there is more annotator uncertainty, and developers should consider the relative risks of omitting a goal, versus attributing an incorrect goal to a user.

In conclusion, we hope that this work provides a basis for robust goal inference evaluation in increasingly complex domains, giving us greater insight into AI behavior and facilitating the development of AI systems that truly act in accordance with our goals.

AI USE STATEMENT

In this work, we used generative AI tools for implementing the web experiment data collection platform, cleaning and reformatting data, visualizing data, writing scripts for making API calls to models, proposing qualitative features of goals and labeling items, and providing feedback on statistical analyses. We have not used generative AI tools for proposing or refining hypotheses, interpreting results, formulating mathematical claims, or writing the manuscript; generating synthetic data sets, proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools for creating figures and identifying existing literature. We have reviewed all AI-assisted work: we manually reviewed papers suggested by an LLM, and manually examined critical statistical analyses for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All human data collection was approved under the IRB of the senior authors' home institution. Annotators were exposed to conversations that potentially involved discussion of high-stakes topics (e.g. relationships, mental health); they were free to leave the experiment at any point and were compensated for their time. We did not collect any personally identifying information about the annotators other than their Prolific ID, which has been removed from the dataset.

This work was motivated by the goal of reducing harm from AI systems, which may arise through incorrect inferences of underlying user goals. However, it is important to note that the goal inference capabilities we examined here could also be employed for potentially detrimental purposes such as manipulation and deception. Therefore, we caution against viewing performance on this benchmark as an overall indicator of model safety.

REPRODUCIBILITY STATEMENT

We have described core details of our data collection and processing pipeline in the main text and Appendix and we have shared our code through the Supplementary Materials portal. Upon deanonymization, we will publicly release all data (with user information removed) and the test harness to facilitate use of the benchmark.

REFERENCES

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56(1): 89–113, 2004. doi: 10.1023/B:MACH.0000033116.57574.95.

Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, et al. Measuring what matters: Construct validity in large language model benchmarks. *Advances in Neural Information Processing Systems*, 38, 2025.

Andrew M Bean, Rebecca Elizabeth Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera-Gómez, Sara Hincapié M, Aruna S Ekanayaka, Lionel Tarassenko, Luc Rocher, et al. Reliability of llms as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, 32(2):609–615, 2026.

Myra Cheng, Isabel Sieh, Humishka Zope, Sunny Yu, Lujain Ibrahim, Aryaman Arora, Jared Moore, Desmond Ong, Dan Jurafsky, and Diyi Yang. Verbalizing llms' assumptions to explain and control sycophancy. *arXiv preprint arXiv:2604.03058*, 2026.

Taha Choukhmane, Tim de Silva, Weidong Lin, and Matthew Akuzawa. AI financial advice: Supply, demand, and life cycle implications. SSRN Working Paper No. 6446286, July 2026. URL `https://ssrn.com/abstract=6446286`.

Fiery Cushman. Rationalization is rational. *Behavioral and Brain Sciences*, 43:e28, 2020.

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.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025.

Chuanyang Jin, Binze Li, Haopeng Xie, Cathy Mengying Fang, Tianjian Li, Shayne Longpre, Hongxiang Gu, Maximillian Chen, and Tianmin Shu. Thoughttrace: Understanding user thoughts in real-world llm interactions. *arXiv preprint arXiv:2605.20087*, 2026.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026.

Rachel Ma, Jingyi Qu, Andreea Bobu, and Dylan Hadfield-Menell. Flexible agent alignment with goal inference from open-ended dialog, 2026. URL `https://arxiv.org/abs/2508.15119`.

Ryan K McBain, Jonathan H Cantor, Joshua Breslau, Melissa Diliberti, Li Ang Zhang, Fang Zhang, Alyssa Burnett, Aaron Kofner, Benjamin Rader, Pat Pataranutaporn, et al. Ai chatbot use and disclosure for mental health among us adolescents and young adults. *JAMA pediatrics*.

Douglas L. Medin, Robert L. Goldstone, and Dedre Gentner. Respects for similarity. *Psychological Review*, 100(2):254–278, 1993.

Richard E. Nisbett and Timothy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977.

Joseph Suh, Ayush Raj, Minwoo Kang, and Serina Chang. Quantifying the utility of user simulators for building collaborative llm assistants. *arXiv preprint arXiv:2605.09808*, 2026.

Jihoon Tack, Philippe Laban, and Jennifer Neville. Llms get lost in evolving user intent. *arXiv preprint arXiv:2607.20734*, 2026.

Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977. doi: 10.1037/0033-295X.84.4.327.

Vijay Veerabadran, Fanyi Xiao, Nitin Kamra, Pedro Matias, Joy Chen, Caley Drooff, Brett Roads, Riley J Williams, Ethan Henderson, Xuanyi Zhao, et al. Benchmarking egocentric multimodal goal inference for assistive wearable agents. *Advances in Neural Information Processing Systems*, 38, 2025.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. In *International Conference on Learning Representations*, volume 2025, pp. 15888–15931, 2025.

Xuhui Zhou, Valerie Chen, Zora Zhiruo Wang, Graham Neubig, Maarten Sap, and Xingyao Wang. Tom-swe: User mental modeling for software engineering agents. *arXiv preprint arXiv:2510.21903*, 2025.

Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, et al. Mind the sim2real gap in user simulation for agentic tasks. *arXiv preprint arXiv:2603.11245*, 2026.

A Appendix

A.1 Additional details for goal elicitation

The ThoughtTrace dataset contained a disproportionate number of conversations related to trip planning (presumably because trip planning appeared as an example task in the onboarding procedure

for that dataset). Therefore, when filtering for high-stakes conversations, we explicitly excluded conversations that discussed trip or vacation planning. For the 30 additional conversations we sampled that were not identified by our classifier as discussing a high-stakes topic but were still included among the set to be annotated, we fixed sampling so that no more than 10 among that set could relate to trip planning.

Annotators could label up to 5 conversations; they had the option to leave at any point before and earn compensation for the conversations they had completed. Annotators were randomly assigned to conversations using an algorithm that had a "budget" (e.g. ensured that a minimum number of annotators were assigned to each given conversation). Annotators earned a base of $12.00 for fully annotating at least one conversation. They were paid an additional $1.00 per completed conversation, and a bonus of $0.10 per additional goal beyond the minimum single likely goal, for up to 4 additional goals per user turn. Annotators were also told that they would be awarded an additional $1.50 per conversation as a quality bonus if their annotations were judged to be in the top 25% of submissions. The median pay (before the quality bonus) was approximately $13 / hour.

Goal data was collected across 2 sessions. To reduce the fractions of annotations that deviated from the task guidelines, we made minor modifications to the onboarding flow and interface for the second session, most notably prepending the text "The user's goal is to:" before each text box where annotators recorded goals. This was to encourage annotators to write from a third-person perspective. After this intervention, the rate of malformed goals dropped to 11.1%, down from 24.4% in the first session.

Malformed entries were disproportionately produced by a relatively small set of annotators; annotators in the top 25% of volume of malformed goals per annotator produced 76% of all malformed goals across the dataset. We removed from the dataset *all* entries that came from annotators in this population, including entries that were not labeled as malformed, under the assumption that these annotators produced systematically lower-quality data. For the remaining 75% of participants, we only dropped entries that were labeled as malformed, maintaining the rest of the data.

A.2 Additional details for automated benchmarking

A.2.1 Human similarity judgment data collection procedure

Annotators were paid a base of $4.50 per conversation for the human-human similarity judgments and $4.00 per conversation for the human-AI similarity judgments (there were fewer goal pairs in the latter setting). They were told they could also earn a "spot check" bonus; they were told that their answers would be assessed against an existing ground truth and if they selected the same option as the ground truth they could earn a bonus of $0.25 per spot check. Finally, they were told that their judgments would be compared against the rest of the human annotators, and for every judgment that aligned with the majority consensus, they would earn an additional $0.05. The median pay (before the majority consensus bonus) was approximately $17 / hour.

A.2.2 Correlation clustering algorithm

In correlation clustering, our goal is to find the clustering assignment π for all human-written goals on a turn $\{g_1, g_2, ..., g_n\}$, that minimizes the following cost function:

$$C(\pi) = \Sigma_{i<j}\Big(\mathbf{1}[\pi(g_i) = \pi(g_j)](3 - \text{sim}(g_i, g_j)) + \mathbf{1}[\pi(g_i) \neq \pi(g_j)]\text{sim}(g_i, g_j)\Big).$$

Intuitively, the cost is low when goals with high similarity ratings (2 or 3) are assigned to the same cluster, and when goals with low similarity ratings (0 or 1) are assigned to different clusters. Because the number of vertices (goals) on each turn is relatively small, the optimization can be solved exactly via integer programming.

A.2.3 Comparison of judge models

Overall results are shown in Table 1. Breakdown by test split for each judge is shown in Tables 2 - 5. All models (including the model that was shown judgments from human-GPT pairs) were modestly better at recovering human-Gemini pairs.

Judge	Cohen's κ	Accuracy	Precision	Recall
Human baseline	0.45	0.759	0.646	0.618
Human–human	0.50	0.779	0.660	0.676
Human–GPT	0.50	0.782	0.672	0.655
Human–Gemini	0.50	0.780	0.664	0.669

Table 1: Agreement between similarity judge and held-out samples under k-fold validation. Judges were evaluated on an approximately equal number of turns from the human-GPT distribution and the human-Gemini distribution.

Evaluation pairs	Cohen's κ	Accuracy	Precision	Recall
Human–GPT	0.47	0.763	0.651	0.638
Human–Gemini	0.44	0.754	0.640	0.597

Table 2: Human baseline, broken down by test set.

Evaluation pairs	Cohen's κ	Accuracy	Precision	Recall
Human–GPT	0.46	0.754	0.625	0.671
Human–Gemini	0.55	0.806	0.704	0.682

Table 3: Human–human judge (provided with in-context examples of human similarity judgments from human-human pairs), broken down by test set.

Evaluation pairs	Cohen's κ	Accuracy	Precision	Recall
Human–GPT	0.49	0.771	0.654	0.676
Human–Gemini	0.51	0.795	0.696	0.630

Table 4: Human–GPT judge (provided with in-context examples of human similarity judgments from human-GPT pairs), broken down by test set.

Evaluation pairs	Cohen's κ	Accuracy	Precision	Recall
Human–GPT	0.46	0.754	0.623	0.678
Human–Gemini	0.55	0.809	0.720	0.659

Table 5: Human–Gemini judge (provided with in-context examples of human similarity judgments from human-Gemini pairs), broken down by test set.

A.3 Model prompts

The prompt used to generate model goals in the *unbounded* setting is as follows:

```
Your task is to infer the goals of a human user who is interacting with an AI agent.
In particular, you should identify all the plausible goals of the person, which defined
as goals that are consistent with what the person has expressed at that point in the
conversation.
Of the plausible goals, you should select a subset as likely goals — which are goals
that are clear enough that a helpful assistant should act on them.
Here are 3 important points to note about the goals.
```

1. Each goal is not just a rephrase or summary or what the person said. It is a clear articulation of the person's intentions, based on the history of the conversation.
   Another way to think about it: anything you write down should be a reasonable continuation of the prompt "The user's goal is to. . . "
2. Each goal you write down is distinct from one another — they should describe meaningfully different things.
   It's ok if 2 goals are similar or related, but they should not be the same.
3. The goals produced should reflect the user's active goals at this point in the conversation — goals that were previously being pursued but have been resolved or abandoned are distinct.

Here are some examples to give you a sense of reasonable goals. The input is the history of the conversation up until the point of goal inference, and the output is the set of user goals. This example will give you a sense of how user goals may or may not change over the course of a conversation. Note that agent messages above a certain length are summarized; there is a flag to indicate if an agent message is a summary, rather than the raw text.

INPUT:

```
[
  {
    'role': 'user',
    'content': 'How to reduce weight',
    'is_summary': None
  }
]
```

OUTPUT:

```
[
  {
    'goal': 'learn about weight-loss strategies',
    'is_likely': True
  },
  {
    'goal': 'improve general health',
    'is_likely': False
  }
]
```

INPUT:

```
[
  {
    'role': 'user',
    'content': 'How to reduce weight',
    'is_summary': None
  },
  {
    'role': 'agent',
    'content': 'The age-old quest to reduce weight! Here are some effective and
    sustainable ways to help you achieve your weight loss goals:\n\n**Diet:**\n-
    Eat a balanced diet of whole, unprocessed foods.\n- Keep track of your calorie
    intake.\n- Stay hydrated and limit sugary drinks.\n\n**Exercise:**\n- Aim for
    at least 150 minutes of aerobic activity per week.\n- Add strength training to
    build muscle.\n\nRemember, aim for a slow and steady loss of 0.5-1 kg per week
    .',
    'is_summary': False
  },
  {
    'role': 'user',
    'content': 'Is 70kgs right weight',
    'is_summary': None
  }
]
```

```
OUTPUT:
[
  {
    'goal': 'determine if the user is a healthy weight',
    'is_likely': True
  },
  {
    'goal': 'gain affirmation that they are not excessively overweight',
    'is_likely': False
  },
  {
    'goal': "figure out the cutoff healthy weight for an upcoming doctor's
    appointment",
    'is_likely': False
  }
]

INPUT:
[
  {
    'role': 'user',
    'content': 'How to reduce weight',
    'is_summary': None
  },
  {
    'role': 'agent',
    'content': 'The age-old quest to reduce weight! Here are some effective and
    sustainable ways to help you achieve your weight loss goals:\n\n**Diet:**\n-
    Eat a balanced diet of whole, unprocessed foods.\n- Keep track of your calorie
    intake.\n- Stay hydrated and limit sugary drinks.\n\n**Exercise:**\n- Aim for
    at least 150 minutes of aerobic activity per week.\n- Add strength training to
    build muscle.\n\nRemember, aim for a slow and steady loss of 0.5-1 kg per week
    .',
    'is_summary': False
  },
  {
    'role': 'user',
    'content': 'Is 70kgs right weight',
    'is_summary': None
  },
  {
    'role': 'agent',
    'content': 'Explains that whether 70 kg is healthy depends on height, body
    composition, age, and activity; introduces BMI (18.5–24.9 as normal) and notes
    70 kg often falls in a healthy range around 170–175 cm, while recommending
    personalized medical guidance.',
    'is_summary': True
  },
  {
    'role': 'user',
    'content': 'Can i eat sweet staff',
    'is_summary': None
  }
]
OUTPUT:
[
  {
    'goal': 'come up with diet plan',
    'is_likely': True
  },
  {
```

```
    'goal': 'get ideas for healthier sweet replacements',
    'is_likely': False
  }
]

You should only produce the OUTPUT list for the provided INPUT. Do not produce any
other text; do not provide any reasoning or explanation.
```

The only difference for the *bounded* prompt was the inclusion of this line, right before the final sentence: "It is very important that you produce exactly 3 goals - any of them can be marked as likely."

A.4 STATISTICAL ANALYSES

All statistical analyses were conducted using the `statsmodels` library. We report the structure of the analyses using `lme4` notation for mixed-effects models.

For the unbounded goals, we evaluated if the model type predicted the number of goals produced on a turn, with a random effect for the turn: `num_goals` $\sim$ `model + (1 | turn_id)`. We fixed Opus as the reference model; coefficients, confidence intervals and p values are shown in Table 6.

For unbounded goals, we tested if the number of goals produced by a model predicted the number of human clusters: `num_human_clusters` $\sim$ `num_model_goals` $\times$ `model` (full results shown in Table 7).

For both the unbounded and bounded goals, we tested if model type predicted precision/weighted recall, with a random effect for the turn: `metric` $\sim$ `model + (1 | turn_id)`, for all combinations of models. We applied a Bonferroni correction for the 105 comparisons, so results were considered significant at $\alpha = 0.05/105 \approx 0.0005$ (full results shown in Figure 6).

For bounded goals, we tested if the presence of a goal feature (for *human* goal features) predicted the percent of models that successfully recovered that goal, for each of the 20 human goal features: `percent_recovered` $\sim$ `feature`. We applied a Bonferroni correction for the 20 comparisons, so results were considered significant at $\alpha = 0.05/20 = 0.0025$ (results are shown in Figure 5). We also conducted a more granular analysis, testing if the goal feature predicted whether an individual model recovered a goal (the outcome variable could take on a value of 0 or 1): `individual_model_recovered` $\sim$ `feature` (again with a Bonferroni correction for 20 comparisons, within models; Figure 7).

For bounded goals, we tested if the presence of a goal feature (for *model* goal features) predicted if that goal was contained in the human clusters, for each of the 20 model goal features, aggregating across *all* model-generated goals: `goal_in_human_data` $\sim$ `feature` (with a correction as for the previous analysis; Figure 8). We also evaluated if feature type predicted representation of that goal in the human data, examining individual models (Figure 9).

For bounded goals, we tested if the number of human clusters predicted precision/weighted recall on that turn, with a random effect for the model and for the turn: `metric` $\sim$ `num_human_clusters + (1 | model_id) + (1 | turn_id)`.

For bounded goals, we tested if the message turn number predicted precision/weighted recall on that turn, controlling for the number of clusters in the ground truth: `metric` $\sim$ `turn_index + C(num_human_clusters) + (1 | model_id) + (1 | turn_id)`. Note that `num_human_clusters` is treated as a categorical variable; the justification is that casting it as a continuous variable would bake in the assumption that the effect of increasing the cluster count by 1 results in equivalent changes from 1 to 2, 2 to 3, etc., which may not be the case.

Table 6: Outputs from $\text{num_model_goals} \sim \text{model} + (1 \mid \text{turn})$. Opus is the reference model.

Model	Mean	β	95% CI	p
Fable 5.1	4.957	-0.586	$[-0.653, -0.520]$	< 0.001
Opus 5	5.544	0	—	—
Sonnet 5	4.325	-1.219	$[-1.286, -1.152]$	< 0.001
Gemini Flash	3.420	-2.124	$[-2.190, -2.057]$	< 0.001
Kimi K3	3.424	-2.120	$[-2.186, -2.053]$	< 0.001
GPT-5.6 Luna	2.983	-2.561	$[-2.627, -2.494]$	< 0.001
GPT-5.6 Sol	2.893	-2.651	$[-2.717, -2.584]$	< 0.001
GPT-5.6 Terra	2.924	-2.620	$[-2.686, -2.553]$	< 0.001
GPT-6 Astra	4.049	-1.495	$[-1.561, -1.428]$	< 0.001
Qwen Flash	4.509	-1.035	$[-1.101, -0.968]$	< 0.001
Qwen Max	3.660	-1.884	$[-1.951, -1.817]$	< 0.001
Inkling	2.988	-2.555	$[-2.622, -2.489]$	< 0.001
Grok 4.6	3.050	-2.494	$[-2.560, -2.427]$	< 0.001
GLM 5.3	3.629	-1.915	$[-1.982, -1.848]$	< 0.001
GLM Flash	3.714	-1.830	$[-1.897, -1.763]$	< 0.001

Table 7: Outputs from $\text{num_human_clusters} \sim \text{num_model_goals} \times \text{model}$. Bold indicates models with a positive and significant slope.

Model	β	95% CI	p
Fable 5.1	$+0.022$	$[-0.089, +0.132]$	0.702
Opus 5	-0.049	$[-0.158, +0.060]$	0.375
Sonnet 5	-0.035	$[-0.145, +0.075]$	0.531
Gemini Flash	**+0.344**	$[+0.153, +0.535]$	< 0.001
Kimi K3	$+0.128$	$[-0.016, +0.272]$	0.082
GPT-5.6 Luna	-0.015	$[-0.119, +0.089]$	0.780
GPT-5.6 Sol	$+0.032$	$[-0.093, +0.157]$	0.615
GPT-5.6 Terra	-0.080	$[-0.192, +0.031]$	0.157
GPT-6 Astra	**+0.149**	**[+0.061, +0.238]**	**<0.001**
Qwen Flash	-0.030	$[-0.107, +0.046]$	0.441
Qwen Max	$+0.051$	$[-0.069, +0.170]$	0.405
Inkling	$+0.097$	$[-0.100, +0.293]$	0.334
Grok 4.6	-0.070	$[-0.233, +0.093]$	0.397
GLM 5.3	-0.127	$[-0.253, -0.000]$	0.049
GLM Flash	-0.137	$[-0.244, -0.029]$	0.013



Figure 6: Pairwise comparisons of model performance (precision and weighted recall), in the unbounded and bounded conditions. Green cells indicate that the row model was significantly better than the column model; red cells indicate that the row model was significantly worse than the column model; white indicates no significant difference.

A.5 Additional details for missed goal analysis

A.5.1 Analysis of human-written goals

All feature descriptions for *human-written* goals are below:

immediate_task_completion: The goal focuses on completing a specific, concrete task or obtaining a particular deliverable directly related to the user's most recent message, such as getting a link, creating a document, finding a recipe, or receiving a recommendation.

broad_life_outcome: The goal describes a general, long-term life outcome, aspiration, or end state that extends well beyond the immediate conversation, such as overall fitness, career success, or general life improvement.

underlying_motivation_attribution: The goal attributes a deeper motivation, reason, or contextual factor behind the user's surface-level request, such as needing a skill for a job, relocating, or pursuing a personal passion.

speculative_uncertain: The goal is hedged with uncertainty markers such as 'may', 'might', 'seems', 'possibly', or 'probably', indicating the annotator is guessing about the user's intentions rather than confidently stating them.

emotional_support: The goal concerns managing the user's feelings, seeking comfort, reassurance, validation, or emotional well-being, such as wanting to feel less anxious, more confident, or less overwhelmed.

social_interaction: The goal involves interpersonal communication, relationship management, social courtesies, or how the user relates to other people (including the AI), such as thanking, apologizing, maintaining peace, or improving communication with someone.

financial_cost: The goal involves saving money, finding affordable or cheap options, budget management, or cost considerations as a primary concern.

avoidance_prevention: The goal is framed as avoiding, preventing, or eliminating something negative rather than achieving something positive, such as avoiding waste, preventing failure, or not exceeding a limit.

format_preference: The goal specifies how information should be presented, structured, or formatted, such as requesting shorter answers, bullet points, tables, step-by-step instructions, or a specific level of detail.

verification_accuracy: The goal involves fact-checking, confirming correctness, validating information, or seeking reassurance that something is accurate, true, or reliable.

skill_building: The goal involves developing, improving, or mastering a specific skill, ability, or competence, such as learning a language, improving communication, or building a technical capability.

decision_selection: The goal involves choosing between options, making a decision, or narrowing down alternatives, such as deciding which product to buy, whether to attend an event, or which approach to take.

comparison_analysis: The goal involves comparing multiple alternatives, evaluating trade-offs, or analyzing differences between options to understand which is better or more suitable.

habit_formation: The goal involves establishing recurring practices, routines, or long-term behavioral changes, such as daily practice, consistent study habits, or maintaining a schedule over time.

meta_conversation: The goal concerns the conversation dynamics, the AI's behavior or capabilities, or interaction management rather than the substantive topic, such as testing the AI, adjusting response style, or ending the conversation.

constraint_satisfaction: The goal includes specific limitations, requirements, or conditions that must be met, such as time limits, dietary restrictions, technical requirements, or situational constraints.

health_wellness: The goal concerns physical health, mental health, fitness, medical concerns, sleep, nutrition, or general well-being as a primary focus.

knowledge_acquisition: The goal is primarily about learning new information, understanding concepts, or gaining knowledge, such as understanding how something works, learning why something happens, or exploring a topic.

future_planning: The goal extends beyond the current conversation to future actions, continued engagement with the AI, or long-term plans, such as returning for more help, implementing advice later, or maintaining an ongoing relationship with the assistant.

self_evaluation: The goal involves assessing one's own performance, knowledge, abilities, readiness, or personal situation, such as testing one's understanding, evaluating one's progress, or reflecting on one's own behavior.

	opus-5	sonnet-5	fable-5.1	gemini-flash	kimi-k3	gpt-5.6-luna	gpt-5.6-sol	gpt-5.6-terra	gpt-6-astra	qwen-flash	qwen-max	inkling	grok-4.6	glm-5.3	glm-flash
immediate task completion	+0.201 p<0.001	+0.192 p<0.001	+0.181 p<0.001	+0.231 p<0.001	+0.218 p<0.001	+0.249 p<0.001	+0.219 p<0.001	+0.226 p<0.001	+0.211 p<0.001	+0.234 p<0.001	+0.184 p<0.001	+0.212 p<0.001	+0.239 p<0.001	+0.219 p<0.001	+0.241 p<0.001
constraint satisfaction	+0.154 p<0.001	+0.147 p<0.001	+0.155 p<0.001	+0.161 p<0.001	+0.170 p<0.001	+0.198 p<0.001	+0.196 p<0.001	+0.184 p<0.001	+0.154 p<0.001	+0.156 p<0.001	+0.141 p<0.001	+0.166 p<0.001	+0.154 p<0.001	+0.165 p<0.001	+0.167 p<0.001
format preference	+0.175 p<0.001	+0.156 p=0.002	+0.113 p=0.027	+0.103 p=0.126	+0.163 p<0.001	+0.131 p=0.010	+0.140 p=0.025	+0.182 p<0.001	+0.099 p=0.480	+0.133 p=0.024	+0.187 p<0.001	+0.180 p<0.001	+0.142 p<0.001	+0.208 p<0.001	+0.183 p<0.001
comparison analysis	+0.109 p=0.009	+0.123 p=0.002	+0.089 p=0.142	+0.118 p=0.011	+0.102 p=0.035	+0.163 p<0.001	+0.162 p<0.001	+0.162 p<0.001	+0.152 p<0.001	+0.163 p<0.001	+0.107 p=0.025	+0.105 p=0.021	+0.120 p=0.022	+0.169 p<0.001	+0.145 p<0.001
decision selection	+0.112 p=0.003	+0.126 p<0.001	+0.107 p=0.006	+0.130 p<0.001	+0.114 p<0.001	+0.126 p<0.001	+0.136 p<0.001	+0.143 p<0.001	+0.109 p=0.003	+0.139 p<0.001	+0.115 p<0.001	+0.109 p=0.010	+0.131 p<0.001	+0.147 p<0.001	+0.162 p<0.001
habit formation	+0.094 p=0.786	+0.115 p=0.141	+0.030 p=1.000	+0.047 p=1.000	+0.088 p=0.331	+0.120 p=0.041	+0.090 p=0.668	+0.068 p=1.000	+0.052 p=1.000	+0.075 p=1.000	+0.053 p=1.000	+0.098 p=0.353	+0.033 p=1.000	+0.131 p=0.002	+0.097 p=0.297
financial cost	+0.074 p=0.401	+0.044 p=1.000	+0.074 p=0.263	+0.058 p=0.889	+0.032 p=1.000	+0.087 p=0.147	+0.057 p=1.000	+0.058 p=1.000	+0.068 p=0.763	+0.043 p=1.000	+0.021 p=1.000	+0.074 p=0.464	+0.023 p=1.000	+0.065 p=1.000	+0.074 p=0.300
verification accuracy	+0.121 p=0.029	+0.033 p=1.000	+0.036 p=1.000	+0.064 p=1.000	+0.071 p=1.000	-0.015 p=1.000	+0.063 p=1.000	+0.082 p=1.000	+0.045 p=1.000	+0.048 p=1.000	+0.065 p=1.000	+0.057 p=1.000	+0.034 p=1.000	-0.005 p=1.000	+0.075 p=1.000
skill building	+0.026 p=1.000	+0.050 p=1.000	+0.020 p=1.000	+0.004 p=1.000	+0.016 p=1.000	+0.082 p=0.867	+0.033 p=1.000	+0.061 p=1.000	+0.054 p=1.000	+0.036 p=1.000	+0.036 p=1.000	+0.076 p=0.422	+0.046 p=1.000	+0.063 p=1.000	+0.037 p=1.000
knowledge acquisition	-0.006 p=1.000	+0.062 p=0.285	-0.029 p=1.000	+0.044 p=0.794	+0.034 p=1.000	+0.054 p=0.445	+0.035 p=1.000	+0.049 p=1.000	-0.001 p=1.000	+0.053 p=0.374	-0.013 p=1.000	+0.022 p=1.000	+0.039 p=1.000	+0.055 p=0.363	+0.049 p=0.717
avoidance prevention	+0.008 p=1.000	-0.007 p=1.000	+0.013 p=1.000	-0.006 p=1.000	+0.022 p=1.000	+0.027 p=1.000	+0.013 p=1.000	+0.047 p=1.000	+0.031 p=1.000	+0.009 p=1.000	+0.044 p=1.000	+0.002 p=1.000	-0.031 p=1.000	+0.038 p=1.000	+0.010 p=1.000
future planning	+0.028 p=1.000	+0.024 p=1.000	+0.004 p=1.000	-0.029 p=1.000	-0.011 p=1.000	+0.007 p=1.000	-0.029 p=1.000	+0.014 p=1.000	-0.030 p=1.000	-0.039 p=1.000	-0.043 p=1.000	+0.046 p=1.000	+0.004 p=1.000	+0.047 p=1.000	+0.002 p=1.000
self evaluation	+0.010 p=1.000	+0.002 p=1.000	-0.072 p=1.000	-0.076 p=1.000	-0.004 p=1.000	-0.049 p=1.000	-0.006 p=1.000	+0.058 p=1.000	+0.033 p=1.000	-0.017 p=1.000	-0.066 p=1.000	-0.029 p=1.000	-0.064 p=1.000	-0.055 p=1.000	+0.006 p=1.000
meta conversation	+0.002 p=1.000	-0.060 p=1.000	+0.051 p=1.000	+0.034 p=1.000	+0.010 p=1.000	-0.171 p<0.001	-0.000 p=1.000	-0.160 p<0.001	+0.040 p=1.000	-0.011 p=1.000	+0.044 p=1.000	-0.056 p=1.000	-0.037 p=1.000	-0.105 p=0.071	-0.064 p=1.000
health wellness	-0.034 p=1.000	-0.024 p=1.000	-0.062 p=0.993	-0.052 p=1.000	-0.024 p=1.000	+0.015 p=1.000	-0.036 p=1.000	-0.012 p=1.000	-0.042 p=1.000	-0.025 p=1.000	-0.064 p=0.696	-0.039 p=1.000	-0.057 p=1.000	+0.012 p=1.000	-0.065 p=0.647
emotional support	-0.013 p=1.000	-0.046 p=1.000	-0.055 p=1.000	-0.112 p=0.021	-0.012 p=1.000	-0.020 p=1.000	-0.063 p=0.850	-0.032 p=1.000	-0.044 p=1.000	-0.009 p=1.000	+0.011 p=1.000	-0.029 p=1.000	-0.070 p=0.978	-0.025 p=1.000	-0.024 p=1.000
social interaction	-0.067 p=0.329	-0.132 p=0.003	+0.004 p=1.000	-0.058 p=0.378	-0.024 p=1.000	-0.086 p=0.184	-0.035 p=1.000	-0.119 p=0.030	+0.043 p=1.000	-0.027 p=1.000	+0.003 p=1.000	-0.067 p=0.925	-0.051 p=1.000	-0.127 p=0.014	-0.072 p=0.284
broad life outcome	-0.058 p=1.000	-0.053 p=1.000	-0.090 p=0.513	-0.095 p=0.087	-0.066 p=0.771	-0.051 p=1.000	-0.094 p=0.230	-0.019 p=1.000	-0.072 p=0.731	-0.086 p=0.415	-0.091 p=0.194	-0.010 p=1.000	-0.087 p=0.283	-0.014 p=1.000	-0.059 p=1.000
underlying motivation attribution	-0.104 p<0.001	-0.144 p<0.001	-0.125 p<0.001	-0.154 p<0.001	-0.159 p<0.001	-0.146 p<0.001	-0.178 p<0.001	-0.111 p<0.001	-0.159 p<0.001	-0.147 p<0.001	-0.164 p<0.001	-0.103 p=0.003	-0.134 p<0.001	-0.144 p<0.001	-0.133 p<0.001
speculative uncertain	-0.335 p<0.001	-0.337 p<0.001	-0.356 p<0.001	-0.368 p<0.001	-0.391 p<0.001	-0.384 p<0.001	-0.395 p<0.001	-0.358 p<0.001	-0.375 p<0.001	-0.405 p<0.001	-0.377 p<0.001	-0.355 p<0.001	-0.395 p<0.001	-0.356 p<0.001	-0.360 p<0.001

Figure 7: Relationship between human goal feature and recovery, broken down by model. Each row represents one of 20 binary features for human-written goals. Each cell is the coefficient of the regression; green represents a positive, significant relationship, red indicates a negative, significant relationship, white represents no significant difference.

A.5.2 Analysis of model-generated goals

Full descriptions of the model goal features are below:

restates_immediate_request: The goal directly addresses or closely paraphrases the specific request, question, or statement made in the user's most recent message — i.e., it is the literal, surface-level reading of what the user just asked for or said (e.g., 'find out the price of the recommended scooter' when the user just asked for the price).

broader_project_goal: The goal frames the user's aim in terms of the larger, ongoing task or project that the conversation has been serving, going beyond the content of the latest message alone (e.g., 'build a personalized English study plan' when the immediate question is only about which skill to focus on first).

longterm_outcome_or_habit: The goal refers to an enduring objective, capability, or sustained pattern of behavior that extends well beyond the current conversation — such as overall language proficiency, health, financial security, career success, or building a consistent daily habit.

emotional_support_seeking: The goal centers on receiving empathy, reassurance, validation, comfort, encouragement, or emotional support from the assistant (e.g., being comforted, feeling less alone, or being reassured that one's feelings or decisions are normal and acceptable).

self_emotion_regulation: The goal involves the user managing, reducing, or overcoming their own internal emotional or cognitive state — such as anxiety, fear, stress, self-doubt, overthinking, frustration, or the need to calm down — as an outcome in itself.

social_companionship: The goal is casual, friendly social interaction, rapport-building, small talk, or companionship with the assistant for its own sake, rather than completing a task or obtaining information.

response_format_control: The goal concerns how the assistant communicates or delivers its help — e.g., response length or brevity, simplicity of language, formatting, tone, pacing (such as going step-by-step or one point at a time), or presenting answers as concise key points.

conversation_closure_gratitude: The goal involves wrapping up, ending, or politely concluding the conversation — including expressing thanks or appreciation, signaling satisfaction, or indicating that no further help is needed right now.

future_engagement_return: The goal involves returning to the assistant later, maintaining an ongoing relationship, reporting back on progress or outcomes, or keeping the option open for future help and follow-up.

assistant_availability_check: The goal involves confirming or testing the assistant's availability, responsiveness, or capabilities — e.g., checking the AI is online and working before making a real request, or probing what it can do.

implementation_follow_through: The goal involves acting on, executing, or putting into practice the advice, plans, decisions, or recommendations already received during the conversation (e.g., starting the routine, making the purchase, contacting someone, following the steps).

information_learning: The goal is primarily to learn, understand, or obtain factual information or knowledge — how something works, what something is, why something happens, or gaining understanding of a topic, mechanism, or process.

recommendation_option_seeking: The goal involves obtaining suggestions, recommendations, ideas, advice, or a menu of options or resources to consider (e.g., recipe ideas, product picks, destination suggestions, resource lists, activity options).

decision_making: The goal involves choosing between alternatives or making a decision or commitment — e.g., deciding between options, whether to do something, or when/how to act on a choice.

plan_creation: The goal involves designing or obtaining a structured plan, schedule, routine, itinerary, roadmap, or organized step-by-step approach to follow.

evaluate_compare_verify: The goal involves comparing, evaluating, assessing, ranking, or verifying — e.g., comparing options or prices, judging suitability, quality, or legitimacy, verifying a claim or deal, or assessing one's own situation, level, or progress.

inferred_underlying_need: The goal attributes to the user a deeper, unstated motivation, concern, or need that goes beyond the literal request — such as reducing overwhelm or decision fatigue, avoiding judgment or embarrassment, gaining confidence, saving money, time, or effort, or protecting one's dignity or peace of mind.

constraint_satisfaction: The goal explicitly incorporates specific constraints or requirements into the aim — such as a budget, time limit, location, dietary or health restriction, skill level, or equipment that the solution must satisfy.

third_party_benefit: The goal centers on the wellbeing, feelings, needs, or preferences of another person or being — such as a friend, family member, partner, child, or pet — rather than (or in addition to) the user themselves.

artifact_deliverable: The goal involves obtaining a concrete, tangible deliverable that can be saved, used, or sent — such as a list, chart, checklist, template, tracker, drafted message, email, script, journal entry, or designed item.

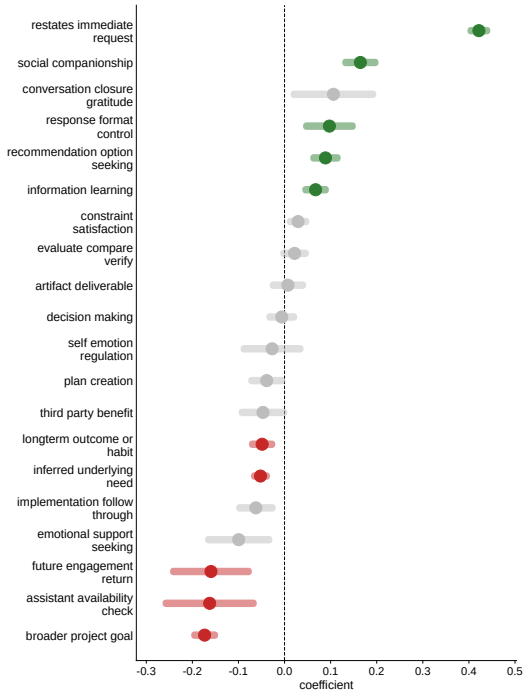


Figure 8: Relationship between model goal types and representation in the human data. Green indicates a positive, significant relationship between the presence of the feature and its representation in the human data. Red indicates a negative, significant relationship. Gray indicates no significant difference. Bars show 95% confidence intervals.

	opus-5	sonnet-5	fable-5.1	gemini-flash	kimi-k3	gpt-5.6-luna	gpt-5.6-sol	gpt-5.6-terra	gpt-6-astra	qwen-flash	qwen-max	inkling	grok-4.6	glm-5.3	glm-flash
restates immediate request	+0.466 p<0.001	+0.514 p<0.001	+0.426 p<0.001	+0.421 p<0.001	+0.396 p<0.001	+0.346 p<0.001	+0.335 p<0.001	+0.406 p<0.001	+0.378 p<0.001	+0.408 p<0.001	+0.389 p<0.001	+0.457 p<0.001	+0.460 p<0.001	+0.468 p<0.001	+0.437 p<0.001
social companionship	+0.131 p=0.029	+0.220 p<0.001	+0.140 p=0.058	+0.178 p<0.001	+0.205 p<0.001	+0.097 p=0.098	+0.119 p=0.014	+0.240 p<0.001	+0.119 p=0.005	+0.107 p=0.257	+0.084 p=1.000	+0.231 p<0.001	+0.154 p=0.008	+0.236 p<0.001	+0.254 p<0.001
conversation closure gratitude	+0.165 p=0.218	+0.273 p<0.001	+0.125 p=0.540	+0.161 p=0.034	+0.060 p=1.000	+0.001 p=1.000	+0.052 p=1.000	+0.102 p=1.000	+0.116 p=0.888	+0.060 p=1.000	+0.045 p=1.000	+0.027 p=1.000	+0.120 p=0.788	+0.245 p=0.022	+0.090 p=1.000
response format control	+0.182 p<0.001	+0.054 p=1.000	+0.122 p=0.013	+0.173 p<0.001	+0.049 p=1.000	+0.092 p=0.433	+0.055 p=1.000	+0.163 p=0.040	+0.104 p=0.064	+0.020 p=1.000	+0.043 p=1.000	+0.086 p=1.000	+0.150 p=0.016	+0.110 p=0.368	+0.153 p=0.120
recommendation option seeking	+0.100 p=0.001	+0.092 p=0.003	+0.108 p=0.001	+0.028 p=1.000	+0.106 p<0.001	+0.081 p=0.008	+0.070 p=0.138	+0.108 p<0.001	+0.071 p=0.060	+0.095 p=0.001	+0.098 p<0.001	+0.111 p<0.001	+0.059 p=0.277	+0.122 p<0.001	+0.102 p=0.007
information learning	+0.100 p=0.002	+0.076 p=0.015	+0.067 p=0.179	+0.082 p=0.003	+0.087 p<0.001	-0.014 p=1.000	+0.032 p=1.000	+0.098 p<0.001	+0.006 p=1.000	+0.049 p=0.873	+0.047 p=0.479	+0.083 p=0.012	+0.112 p<0.001	+0.111 p<0.001	+0.058 p=0.378
constraint satisfaction	+0.048 p=0.194	+0.011 p=1.000	+0.068 p=0.027	+0.059 p=0.038	+0.034 p=1.000	+0.007 p=1.000	+0.005 p=1.000	-0.002 p=1.000	+0.017 p=1.000	+0.028 p=1.000	+0.068 p=0.014	+0.034 p=1.000	+0.019 p=1.000	-0.003 p=1.000	+0.042 p=1.000
evaluate compare verify	-0.013 p=1.000	+0.044 p=1.000	+0.024 p=1.000	+0.015 p=1.000	+0.009 p=1.000	+0.022 p=1.000	-0.005 p=1.000	+0.048 p=0.887	-0.010 p=1.000	+0.002 p=1.000	+0.012 p=1.000	+0.026 p=1.000	+0.081 p=0.124	+0.041 p=1.000	-0.000 p=1.000
artifact deliverable	-0.024 p=1.000	-0.002 p=1.000	+0.024 p=1.000	+0.092 p=0.075	+0.000 p=1.000	+0.030 p=1.000	+0.003 p=1.000	+0.059 p=1.000	+0.073 p=0.182	+0.040 p=1.000	+0.033 p=1.000	-0.021 p=1.000	-0.052 p=1.000	-0.051 p=1.000	-0.004 p=1.000
decision making	-0.027 p=1.000	-0.018 p=1.000	+0.015 p=1.000	+0.000 p=1.000	-0.044 p=1.000	+0.021 p=1.000	-0.018 p=1.000	+0.004 p=1.000	+0.021 p=1.000	-0.005 p=1.000	-0.016 p=1.000	+0.011 p=1.000	+0.015 p=1.000	-0.025 p=1.000	-0.037 p=1.000
self emotion regulation	-0.051 p=1.000	+0.026 p=1.000	-0.019 p=1.000	-0.048 p=1.000	-0.029 p=1.000	-0.049 p=1.000	+0.024 p=1.000	-0.069 p=1.000	-0.023 p=1.000	-0.037 p=1.000	-0.106 p=0.324	-0.007 p=1.000	+0.050 p=1.000	-0.006 p=1.000	-0.015 p=1.000
plan creation	-0.049 p=1.000	-0.094 p=0.025	+0.033 p=1.000	+0.026 p=1.000	-0.035 p=1.000	-0.031 p=1.000	+0.001 p=1.000	-0.030 p=1.000	+0.028 p=1.000	-0.039 p=1.000	-0.050 p=1.000	-0.042 p=1.000	-0.065 p=1.000	-0.111 p=0.014	-0.055 p=1.000
third party benefit	+0.014 p=1.000	-0.057 p=1.000	+0.004 p=1.000	-0.062 p=1.000	+0.016 p=1.000	-0.085 p=0.227	-0.096 p=0.181	-0.025 p=1.000	-0.045 p=1.000	-0.073 p=1.000	-0.043 p=1.000	-0.093 p=1.000	-0.065 p=1.000	-0.037 p=1.000	-0.063 p=1.000
longterm outcome or habit	-0.062 p=0.227	-0.069 p=0.122	-0.035 p=1.000	-0.048 p=0.717	-0.037 p=1.000	-0.033 p=1.000	-0.018 p=1.000	-0.079 p=0.033	-0.060 p=0.238	-0.070 p=0.203	-0.054 p=0.363	-0.021 p=1.000	-0.060 p=0.158	-0.035 p=1.000	-0.058 p=0.517
inferred underlying need	-0.089 p<0.001	-0.016 p=1.000	-0.073 p=0.004	-0.103 p<0.001	-0.000 p=1.000	-0.065 p=0.050	-0.063 p=0.004	-0.045 p=0.539	-0.100 p<0.001	-0.065 p=0.020	-0.070 p=0.010	-0.034 p=1.000	-0.053 p=0.292	+0.011 p=1.000	-0.033 p=1.000
implementation follow through	-0.037 p=1.000	-0.073 p=0.904	-0.021 p=1.000	-0.057 p=1.000	-0.086 p=0.209	-0.076 p=0.162	-0.046 p=1.000	-0.139 p<0.001	-0.059 p=1.000	-0.026 p=1.000	-0.031 p=1.000	-0.094 p=0.025	-0.044 p=1.000	-0.105 p=0.046	-0.021 p=1.000
emotional support seeking	-0.156 p=0.032	-0.011 p=1.000	-0.096 p=0.965	-0.092 p=1.000	-0.088 p=0.550	-0.107 p=0.990	-0.061 p=1.000	-0.076 p=1.000	-0.089 p=1.000	-0.155 p=0.059	-0.182 p<0.001	-0.071 p=1.000	-0.079 p=1.000	-0.103 p=0.401	-0.012 p=1.000
future engagement return	-0.092 p=1.000	-0.027 p=1.000	-0.326 p<0.001	-0.308 p<0.001	-0.146 p=1.000	-0.028 p=1.000	-0.057 p=1.000	-0.041 p=1.000	-0.073 p=1.000	-0.144 p=1.000	-0.130 p=1.000	-0.142 p=0.607	-0.211 p=0.087	-0.236 p=0.027	-0.125 p=1.000
assistant availability check	-0.088 p=1.000	-0.084 p=1.000	-0.256 p=0.016	-0.237 p=0.003	-0.283 p=0.013	+0.097 p=1.000	-0.105 p=1.000	-0.002 p=1.000	-0.251 p=0.044	-0.115 p=1.000	-0.196 p=0.031	-0.199 p=0.560	-0.186 p=0.595	-0.296 p=0.006	-0.287 p=0.001
broader project goal	-0.187 p<0.001	-0.197 p<0.001	-0.166 p<0.001	-0.168 p<0.001	-0.152 p<0.001	-0.168 p<0.001	-0.121 p<0.001	-0.206 p<0.001	-0.133 p<0.001	-0.167 p<0.001	-0.146 p<0.001	-0.197 p<0.001	-0.191 p<0.001	-0.252 p<0.001	-0.172 p<0.001

Figure 9: Relationship between model goal feature and representation of the goal in the human data, broken down by model. Each row represents one of 20 binary features for model-generated goals. Each cell is the coefficient of the regression; green represents a positive, significant relationship, red indicates a negative, significant relationship, white represents no significant difference.