

# Large language models accurately identify decision reasons in verbal reports

Participant-visible method context. Results, discussion, captions, and the target figure are excluded.

**Decision Reasons.** To characterize the decision making of participants from their verbal reports, we defined a structured set of 47 candidate decision reasons. Each reason is an operational construct designed to capture a potential strategy or principle for choosing between risky options. In our framework, a decision reason must meet a minimal discriminability criterion: it must specify a consistent

PNAS

Vol. 123

No. 27

e2526798123

Inference with LLMs. For the initial model comparison, we used both

proprietary (GPT-4o, GPT-4o-mini) and open-weight LLMs. The proprietary models were accessed via the Azure API at the Max Planck Institute for Human Development. Open-weight models (Llama, and Qwen) were run locally on a high-performance computing cluster using Hugging Face (see, ref. 1,

<https://doi.org/10.1073/pnas.2526798123>

9 of 12

for a tutorial) and vLLM (<https://docs.vllm.ai/>). All models were run with default inference settings (e.g., temperature) to assess their typical performance. Critically, all verbal reports were analyzed in their original German without any preprocessing, cleaning, or translation. This included typographical errors, grammatical inconsistencies, and informal phrasing, allowing us to evaluate the LLM's ability to handle raw, naturalistic text. For the main analysis, we selected Qwen3-235B-A22B. This model achieved the highest accuracy in our initial validation (Fig. 2), outperforming all other tested models. Its open-weight and multilingual nature made it an ideal choice for analyzing the German-language reports in a transparent and reproducible manner. The complete codebase and model outputs are available in our project repository.

choices. Because multiple reasons could be identified for a single trial, we used a majority-rule aggregation to derive one predicted choice per trial. First, we mapped choice options to numeric values: Option B was coded as  $-1$ , Option A as  $+1$ , and indifference (used only for reason-based preferences) as  $0$ . Let  $RT,i$  be the set of decision reasons identified for trial  $i$  at threshold  $T$ . This set can be represented as a vector of reason-implied preferences, where each reason  $r \in RT,i$  is uniquely mapped to a single value  $vr \in \{-1, 0, +1\}$  according to its formal rule. Second, the model's predicted choice for trial  $i$  at threshold  $T$  was then calculated as  $\square \square \sum_{r \in RT,i} vr \square$ ,  $\Upsilon T,i = \text{sign} \square$

Downloaded from <https://www.pnas.org> by UNIVERSITAETSBIOTHEK TUEBINGEN on July 9, 2026 from IP address 134.2.76.160.

**Behavioral Experiment.** The study was conducted in person in the laboratories

of the Max Planck Institute for Human Development in Berlin. A total of 86 adults ( $M = 28.4$ ,  $SD = 6.2$ ; 62% female; 50% university degree) participated after giving informed consent. The study was approved by the Institutional Review Board (Ethics Committee) at MPI for Human Development. Participants were compensated with a base payment of 17 euros plus a performance-based bonus (up to  $\pm 3$  euros) determined by playing out one randomly selected choice trial for real money. Participants completed 20 monetary choice problems in a randomized order. The task was presented on a computer using an interactive interface where lottery outcomes and probabilities were initially hidden and could be revealed by mouseover. After making their choice in each trial, a text box appeared with the following prompt: Bitte beschreiben Sie WIE und WARUM Sie diese Entscheidung getroffen

haben. Wir möchten erfahren, mit welchen Schritten Sie Informationen gesammelt haben und welche Gründe Sie für Ihre finale Entscheidung in Betracht gezogen haben im Hinblick auf die Eigenschaften der Lotterien. Geben Sie die Schritte, wenn möglich, in der Reihenfolge an, in der sie stattgefunden haben. (English translation: Please describe HOW and WHY you made this decision. We would like to know what steps you took to gather information, and what reasons you considered for your final decision in relation to the characteristics of the lotteries. If possible, list the steps in the order they occurred.) Due to a partial loss of early source code, the exact prompt phrasing for the first 41 participants is unavailable, but it was effectively equivalent. The correlation between the proportions of trials in which a given reason was identified across sessions was  $r = 0.98$ , with a very small average absolute difference ( $M = 0.003$ ,  $SD = 0.003$ ), indicating that the distributions of identified reasons did not differ between sessions. To validate our choice data, we compared the choice proportions for the 16 canonical problems to those from a large-scale replication study (43). The correlation between the choice proportions was very high ( $r = 0.87$ ), and the median absolute difference was small ( $Me = 0.11$ ), indicating that the addition of a verbal reporting task did not systematically distort decision behavior. We collected a total of 1,720 free-text reports. The reports varied in length, with a median of 27 words ( $Q1 = 16$ ,  $Q3 = 42$ ) and 165 characters ( $Q1 = 96$ ,  $Q3 = 258$ ). The vast majority of participants provided substantive responses suitable for semantic analysis. Choice Prediction. To evaluate whether the LLM-identified decision reasons

could predict participants' choices, we implemented two prediction strategies. First, we used a majority-rule approach to determine the optimal confidence threshold for identifying the presence of a reason (Fig. 3B). Second, we applied a reason-weighted aggregation model to test whether contextual distributions of reasons could generalize to unseen trials in an out-of-sample setting (Fig. 6). Together, these methods assess both the within-sample alignment and the broader predictive utility of our approach. Determining a threshold for reason identification. To determine an appropriate confidence threshold for identifying decision reasons, we tested how well the reasons selected at different levels aligned with participants' actual

10 of 12

<https://doi.org/10.1073/pnas.2526798123>

where  $\text{sign}(\cdot)$  returns +1 if the sum is positive (predicting Option A), -1 if negative (predicting Option B), and 0 in case of a tie. To compute prediction accuracy, we compared the predicted choice  $\hat{Y}_{T,i}$  with the observed choice  $Y_{i,obs}$  :

$Y_i$

=

( 1 if  $\hat{Y}_{T,i} = Y_{i,obs}$  0 if  $\hat{Y}_{T,i} \neq Y_{i,obs}$  .

A tie  $\hat{Y}_{T,i} = 0$  was scored 0.5 to reflect a random guess. Finally, we aggregated accuracy at the participant level by averaging each participant's alignment scores across their 20 trials. We evaluated thresholds  $T \in \{0, 10, 20, \dots, 100\}$ , as the LLM's confidence scores were provided in discrete steps of 10. The resulting distribution of participant-level accuracies is shown as a boxplot for each threshold in Fig. 3B. Out-of-sample predictions. This analysis tested whether the distribution of identified reasons within a specific context (e.g., for a particular problem or participant) could predict choices for new trials from that same context. This assesses if the observed variability in reason use reflects a consistent and generalizable structure. First, we constructed a binary reason-identification matrix  $X \in \{0, 1\}^{n \times R}$ , where  $n = 1,720$  is the number of trials and  $R = 47$  is the number of reasons included in the final analysis set. An entry  $X_{i,r} = 1$  indicates that reason  $r$  was identified for trial  $i$  (i.e., LLM confidence  $\geq 80$ ); otherwise,  $X_{i,r} = 0$ . Context data for each trial included participant ID, participant cluster (Fig. 5A), problem ID, problem type (canonical types, and problem cluster (Fig. 5B). Second, the data were split into training and test sets using an 80–20 stratified split within each participant, resulting in 16 training and 4 test trials per participant per iteration. Let  $D_{train}$  and  $D_{test}$  denote the resulting subsets. For a given context  $C$  (e.g., participants, problems, or clusters) with levels  $\{C_1, \dots, C_J\}$ , we then computed the frequency of each reason within the training trials of level  $C_j$ , defined as  $IC_j = \{i \in D_{train} : C(i) = C_j\}$ . The frequency vector was given by  $\sum_{i \in IC_j} fC_j[r] = X_{i,r}$ ,  $i \in IC_j$

yielding  $fC_j \in \mathbb{N}^R$ , which represents the count of identified reasons for level  $C_j$ . Third, we defined a reason-preference matrix  $R_{pref} \in \{-1, 0, +1\}^{D \times R}$ , where  $D = 20$  was the number of unique choice problems. Each entry  $R_{pref}[d, r]$  specifies the preference implied by reason  $r$  in problem  $d$ : +1 for Option A, -1 for Option B, and 0 if indifferent. For each test trial  $i \in D_{test}$ , we used the corresponding condition level  $C_j(i)$  and problem  $d(i)$  to

compute the predicted choice:  $\hat{Y}_i = \text{sign} \left( \sum_r f_{Cj}(i) [r] \cdot R_{\text{pref}}[d(i), r] \right)$ ,  $r = 1$

with  $\hat{Y}_i = +1$  for Option A,  $-1$  for Option B, and  $0$  for ties. Finally, prediction accuracy was evaluated analogously to the threshold-based analysis: each test trial contributed  $1$  for a correct prediction,  $0$  for an incorrect prediction, and  $0.5$  for a tie. The participant-level mean accuracy was computed with these scores. This entire procedure was repeated for 5,000 random splits.

pnas.org

The resulting distributions of mean participant-level out-of-sample accuracies are shown as boxplots in Fig. 6. Perplexity. To characterize how many reasons effectively contribute to each predictive model's predictions, we calculated the perplexity of the context-specific reason distributions. Perplexity is an information-theoretic measure that captures the effective number of categories in a distribution: it equals  $1$  when all weight is concentrated on a single reason and equals the total number of reasons when weight is spread uniformly. Lower perplexity thus indicates a more concentrated, parsimonious reason distribution. Specifically, within each out-of-sample iteration, the frequency vector  $f_{Cj}$  for a given level  $C_j$  was normalized to form a probability distribution  $p_{Cj}$ , defined as  $p_{Cj}[r] = \frac{f_{Cj}[r]}{\sum_r f_{Cj}[r]}$

$f_{Cj}[r] / \sum_r f_{Cj}[r]$

The perplexity  $PPC_j$  for each profile was then computed as  $PPC_j = 2^{H(p_{Cj})}$

$H(p_{Cj})$

where

$H(p_{Cj}) = -$

$\sum$

$p_{Cj}[r] \log_2 p_{Cj}[r]$

Downloaded from <https://www.pnas.org> by UNIVERSITAETS BIBLIOTHEK TUEBINGEN on July 9, 2026 from IP address 134.2.76.160.

$r = 1$

is the Shannon entropy of the distribution. For each context  $C$  (e.g., problem cluster), we calculated the average perplexity across all its constituent levels  $\{C_1, \dots, C_J\}$ . The values reported in Fig. 6 reflect the mean of these average perplexity scores aggregated over all 5,000 out-of-sample iterations. Prospect theory. To benchmark our reason-based models, we implemented a hierarchical version of prospect theory (PT). Consistent with recent large-scale findings (13), we used the original, noncumulative version. The subjective value  $V(A)$  of a lottery  $A$  with outcomes  $x_i$  and probabilities  $p_i$  is  $V(A) =$

$\sum$

$w(p) = \exp(-(-\ln p)^\alpha)$ . The choice probability was determined by a logistic-softmax function:  $P(A) =$

$\frac{\exp(-\lambda \cdot (V(A) - V(B)))}{1 + \exp(-\lambda \cdot (V(A) - V(B)))}$

where  $\lambda$  is a sensitivity parameter. We estimated the model using a Bayesian hierarchical framework in Stan, with flat priors, obtaining population and individual-level estimates for  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\lambda$ . Out-of-sample predictive accuracy (Fig. 6) was estimated using Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO) via the loo package in R (58). The median posterior parameter estimates (with 95% Bayesian CI; BCI) were  $\alpha = 0.34$ , 95% BCI: [0.23, 0.53];  $\beta = 1.68$ , 95% BCI: [0.98, 2.86];  $\gamma = 0.73$ , 95% BCI: [0.57, 0.88];  $\lambda = 0.41$ , 95% BCI: [0.14, 0.74]. Clustering. To explore variability in reasoning patterns, we applied a dimensionality reduction and clustering pipeline to the reason-by-individual and reason-by-problem matrices. First, we used PaCMAP (45) to embed the high-dimensional matrices into a two-dimensional space. Second, we performed hierarchical clustering

on the resulting embeddings using Ward's linkage criterion. Consistent with the analysis in the main text and our desire for a parsimonious description, we extracted four clusters for participants and three for choice problems. These cluster assignments were used for visualization (Fig. 5) and as a conditioning variable in the out-of-sample prediction analysis (Fig. 6). This analysis was exploratory and intended to illustrate the structure present in the reason distributions.