How do we get to know someone? Diagnostic questions for inferring personal traits

Erik Brockbank*, Tobias Gerstenberg, Judith E. Fan, & Robert D. Hawkins
Department of Psychology, Stanford University, USA
*ebrockbank@stanford.edu

Abstract

When first meeting somebody, we're faced with the challenge of "getting to know them." Why do some questions seem to enable this better than others? In Experiment 1, participants ($N = 185$) evaluated a large bank of conversational questions. We found that questions varied along a reliable latent dimension of interpersonal depth ranging from "small talk" to "deep" questions. In Experiment 2 ($N = 188$), participants answered a subset of these questions along with a number of self-report personality scales. Using a language model to estimate how informative participants' free responses were, we find that individualized personality predictions were more accurate when incorporating free responses; furthermore, responses to deeper questions supported more accurate personality inferences than small talk. Taken together, results suggest not only that responses contained the statistical information necessary to make abstract social inferences, but also that people have accurate intuitions about which conversational topics enable learning about and connecting with others.

Keywords: social learning; personality; question asking; closeness; language models

Introduction

Starting from a young age, humans learn the dance of meeting someone new, whether it's a colleague, a classmate, or the person next to you on a flight. An observer of these exchanges might notice that they typically involve some questions ("Where are you from?") more than others ("How many toes do you have?"). While these conversations respect a range of norms and situational constraints, "getting to know somebody" includes epistemic goals—towards this end, some questions are more useful for learning what others are like. What kinds of questions help us get to know others, and what information about a speaker do the answers contain?

In the current work, we measure the *diagnostic* value of different questions for learning what others are like. In fact, this problem is familiar to psychologists interested in identifying stable differences between people. Prominent theories of human personality propose succinct axes that account for differences between individuals, collapsing the many ways people vary in their actions, beliefs, and motivations onto a more tractable set of latent dimensions (Allport, 1961; Cervone & Pervin, 2022). For instance, a large body of work on the *Big Five* personality traits suggests that many people can be loosely individuated by their expression of just five abstract traits, usually referred to as, "openness," "conscientiousness," "extraversion," "agreeableness," and "neuroticism" (Goldberg, 1993).

To measure what a person is like along these dimensions, psychologists rely on questions designed to be maximally diagnostic. For instance, the Big Five Inventory asks respondents to indicate how strongly they agree with statements such as, "I see myself as someone who is curious about many different things" (John & Srivastava, 1999). The standard Big Five Inventory contains 44 such questions, yet efforts to identify the most diagnostic items have distilled the original inventory down to as few as 10 (Rammstedt & John, 2007; Gosling, Rentfrow, & Swann Jr, 2003). What makes these questions diagnostic is the degree to which individual responses vary and "hang together" across questions, allowing psychologists to reliably map individuals onto the underlying dimensions that best differentiate them. In addition to their psychometric properties, the usefulness of these questions for measuring what others are like has been assessed from a range of perspectives, including their stability across languages and cultures (Steyn & Ndofirepi, 2022; Costa Mastrascusa et al., 2023; Gurven, Von Rueden, Massenkoff, Kaplan, & Lero Vie, 2013) and over the lifespan (Rothbart, Ahadi, & Evans, 2000; Shiner, Soto, & De Fruyt, 2021).

Despite the value of psychological measures such as the Big Five Inventory for differentiating individuals, people are rarely seen with pen and clipboard eliciting sliding scale judgments from others, even in contexts where the goal is to get to know them. Instead, people rely on everyday conversation to acquire a model of what others are like (Aron, Melinat, Aron, Vallone, & Bator, 1997; Mahaphanit, Welker, Schmidt, Chang, & Hawkins, 2024). Yet compared to the scientific measurement of personality, comparably little is known about the process by which we get to know others using natural language—our *intuitive personality psychology*.

The current work seeks to quantify the diagnostic value of everyday questions for learning about others. In Experiment 1, we developed a corpus of open-ended questions ranging from "small talk" to "deep" questions. Participants evaluated these questions on a series of scales related to the question's effectiveness for getting to know others, allowing us to probe people's intuitive model of how different questions might support social inference. In Experiment 2, we then explore the relationship between written responses to questions from Experiment 1 and the same participants' responses to personality scales commonly used in psychological research. We analyze predictions from a large language model to esti-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

mate whether evaluations of the *questions* from Experiment 1 reflect underlying differences in what the *answers* say about a speaker. Overall, the results suggest that people have an intuitive understanding of the interpersonal depth of different questions for getting to know others, and that these judgments are modestly reflected in differences in the information about a speaker that the question's answer contains.

Exp 1: Evaluating question depth

Participants

200 participants were recruited from Prolific in order to obtain roughly 10 evaluations for each of the questions participants evaluated. Two were excluded due to technical errors during data collection. An additional 13 were removed due to failure to respond correctly to one or more attention checks embedded in the task, leaving $N = 185$ participants (*age*: median: 33, range: 18-74; *gender*: 82 female, 99 male, 3 nonbinary; *race*: 107 white, 51 black, 19 Asian, 6 multiracial). All participants were fluent English speakers. Participants were paid $3.75 for an estimated 15 minutes to complete the study (*median completion time*: 13m 7s) following the Stanford University IRB protocol.

Stimuli

Question bank Participants were asked to evaluate questions from a large question bank of open-ended questions (i.e., questions which can only be answered through free response, not by an agreement scale or multiple choice).[1] The question bank contained 235 questions in total, 125 *personal* questions and 110 *small talk* questions (Figure 1). A subset of the questions in each category were taken from the "small talk" and "closeness-generating" questions given to participants in Aron et al. (1997). The remaining questions were generated by the authors and colleagues. When eliciting potential questions, colleagues were instructed to think of questions which they felt would help them better get to know somebody (*personal* questions) or, in contrast, questions which would *not* help them get to know a conversation partner (*small talk* questions). This distinction guided the sampling procedure during the task but was otherwise irrelevant since participants evaluated the questions directly (including with respect to how much or how little they would help in getting to know somebody).

Question evaluation scales Participants evaluated questions from the question bank using nine evaluation scales. All scales were presented as continuous sliders with labels at each endpoint. The scales were selected to capture intuitive features of the questions that might contribute to how well they help people get to know each other (e.g., how *informative* a question is; see complete list in Figure 2A).

[1] All code and materials available at: https://github.com/erik-brockbank/deep_conversations_cogsci2025

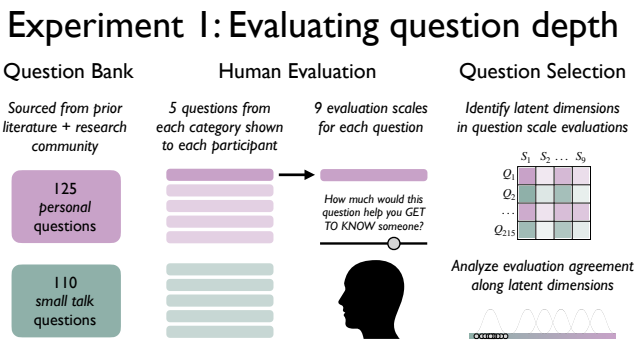

Figure 1: Experiment 1 overview. Participants evaluated questions from a large question bank of *personal* and *small talk* questions using nine distinct evaluation scales.

Procedure

Participants were told they would be evaluating a series of questions that might be asked in conversation and were instructed to assume that they or a conversation partner answered honestly. The 10 questions were selected by sampling five from the *personal* question category and five from the *small talk* question category in the question bank. The order of the questions was shuffled and participants were not told about the category or the origin of the questions. The nine evaluation scales were divided into four blocks that grouped similar scales (e.g., how much participants would expect to *learn* about a conversation partner and how much they would expect a conversation partner to learn about *them*). For each participant, the order of the blocks was shuffled as well as the order of the questions *within* each block. This allowed the order of the scales to vary between participants while always presenting similar scales together.

Participants performed each evaluation one at a time and were not permitted to change their answer once they had submitted it. During each trial, they were shown counters indicating the question number, the total number of questions they would evaluate, and the index of the current evaluation scale alongside the text of the current question and evaluation scale. The text of the evaluation scales were color coded and formatted to streamline visual processing since many of the scales were similar in their wording.

Interleaved among the evaluations, participants completed two attention check trials in which they were shown a novel scale asking them to move the slider to one labeled endpoint (e.g., "Please drag the slider all the way to the end labeled HAMBURGER"). Attention check trials were inserted by first sampling a question index (excluding the first and last) and then sampling a trial index among the evaluations for that question. After completing all evaluation trials, participants were shown a brief post-experiment survey which included demographic questions as well as two free response questions allowing them to report any technical difficulties or feedback.

Results

Questions vary along a latent dimension of depth Overall, evaluations of each question were highly reliable across the nine scales (Cronbach's $\alpha = .946$, 95% CI = $[.941, .951]$). To better understand the relationship between evaluations on each of the scales, we fit a factor analysis model to the responses. The factor analysis allows us to investigate how well participants' evaluations of each question across the nine scales can be approximated by a more compact set of dimensions, each of which is a linear combination of the scales. This approach embodies the idea that insofar as responses on some of the scales "hang together" (e.g., "How *deep* is this question" and "How *personal* is this question" may produce similar evaluations), participants' responses can be described by a smaller set of *latent* variables that combine the evaluation scales (Eisenberg et al., 2019).

We found that 83% of the variance in participants' average evaluations of each question can be accounted for by a single latent factor, while the second and third factors each account for roughly 4% and 3% of the remaining variance to be explained (Figure 2A). All nine scales load fairly uniformly onto this first factor, suggesting that participants treated each of the nine scales as interrogating highly overlapping aspects of the questions. Loadings for the second factor were positive for scales related to how *personal* and *deep* a question is and whether it leads to *closeness*; meanwhile, loadings were negative for scales related to the question's *informativity*, usefulness in *getting to know* someone, and helping to *learn* about a speaker. This suggests that some of the questions may have pulled participants in opposite directions with respect to affective and epistemic features of the question. Taken together, these results are consistent with participants representing each question along a single intuitive dimension related to its *interpersonal depth* and choosing scale evaluations that reflect this underlying dimension.

Observers agree about question depth How were *personal* and *small talk* evaluations distributed along this latent depth scale? We projected participants' responses for each question onto the first factor using the loadings estimated by our factor analysis. This offers insight into the variability *between* questions on the first factor, as well as variability *within* question based on individual evaluations on that same axis. Question evaluations varied along the latent interpersonal depth dimension and exhibited similar agreement among participants at both high and low ends of the scale (see Figure 2B for a sample of questions at either extreme of the scale). We use each question's estimated value on the interpersonal depth dimension and the variance of this estimate to select a subset of questions which were among the highest and lowest in overall question depth and had the highest agreement. These items constitute our most reliable estimate of "small talk" and "personal" questions based on participants' own intuitive judgments about the questions. We use these questions in Experiment 2 to ask what information about a speaker is encoded in their *responses*.

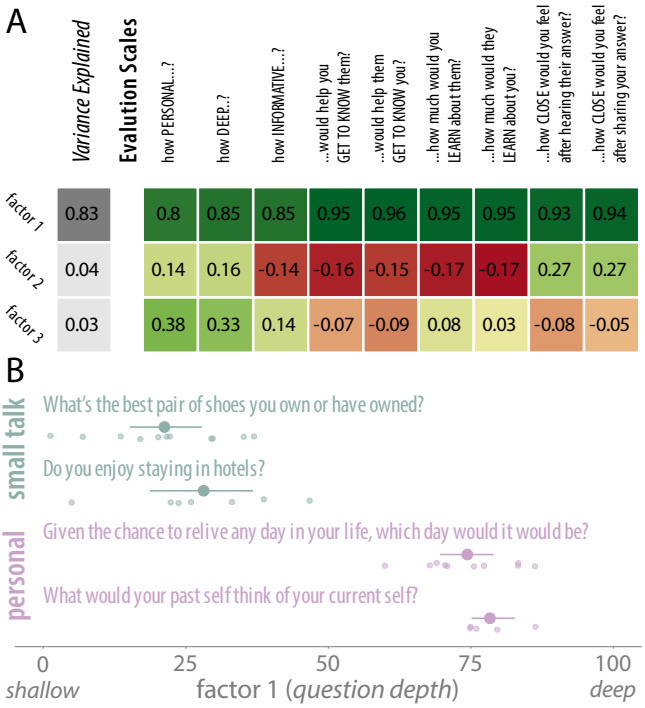

Figure 2: A) Factor loadings for the first three factors of participants' question evaluations (scale text is truncated for demonstration). B) Sample question ratings projected onto factor 1. Error bars indicate 95% bootstrapped confidence intervals. Question text is shortened for demonstration.

Exp 2: What do responses say about speakers?

The results of Experiment 1 suggest that questions vary in how *informative* they are, how well they help us *get to know* others, and a range of similar evaluation scales. These judgments reliably placed questions at different points along a single latent dimension that may reflect a question's perceived level of interpersonal depth. These evaluations provide one estimate of a question's diagnostic value for getting to know others. In Experiment 2, we obtain a second estimate of this value by asking how much *responses* to a question are informative about speakers.

Participants

210 participants were recruited from Prolific. This number was chosen to target roughly 100 answers for each of the free response questions used in the experiment (each participant answered half of the available questions) with additional buffer for attention check failures and other exclusions. Of the 210 participants recruited, 16 were excluded due to technical error during data collection and five were excluded due to failure to respond correctly on one or more of the attention checks. One additional participant was excluded due to clear evidence of using an LLM or other AI solution to answer the free response questions, leaving $N = 188$ participants (*age*: median: 36, range: 19-74; *gender*: 78 female, 100 male, 2

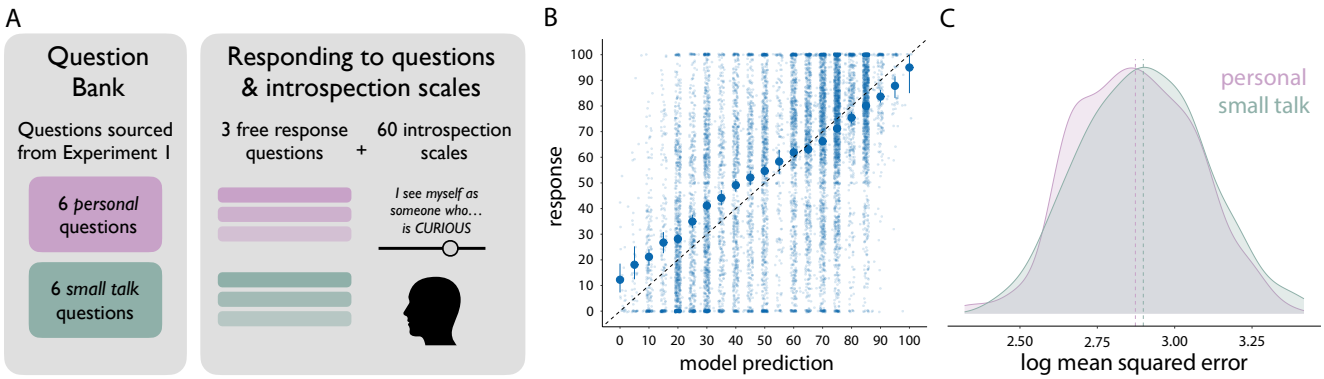

Figure 3: (A) Experiment 2 overview. (B) Relationship between language model predictions and participants' introspective evaluations. (C) Log mean squared error distributions for model predictions using *personal* and *small talk* questions.

non-binary; *race*: 119 white, 39 black, 7 Asian, 10 multiracial). All participants were fluent English speakers. Participants were paid $5 for an estimated 20 minutes to complete the study (median completion time: 23m 12s) following the Stanford University IRB protocol.

Stimuli

Question bank Using results from Experiment 1, we selected six low-depth *small talk* questions and six high-depth *personal* questions whose evaluations were clustered at the extremes of the evaluation scales and which further exhibited high interrater agreement in their evaluations. In addition to these criteria, we sought questions for which participants could easily spend up to two minutes writing a response.

Personality scales In addition to the free response questions, participants were prompted with 60 continuous slider scales, each asking for a distinct subjective evaluation; these personal introspection scales were chosen from three broad categories. First, 44 items came from the Big Five Inventory, a measure of individual variation along five personality dimensions that has been extensively validated in a range of languages and cultures (John & Srivastava, 1999; Costa Mastrascusa et al., 2023; Rammstedt & John, 2007; Steyn & Ndofirepi, 2022). Questions from the inventory primarily take the form "*I see myself as someone who...*" with completions such as "*is generally trusting*". Participants responded on a continuous slider with "Strongly disagree" and "Strongly agree" at either endpoint (Figure 3A).

A second set of introspection scales was drawn from prior research exploring the social inferences people make when seeing unfamiliar faces. The 15 items used in Oosterhof and Todorov (2008) were adapted to match the format of the Big Five questions ("I see myself as someone who..."); two were removed due to overlap with the Big Five and one was removed because it was not relevant to the current task (attractiveness), leaving 12 items which have been shown to encompass inferences made about others when viewing their faces.

Finally, we included four items that capture distinct aspects of personal identity but are not present in the Big Five or in prior face processing work: two questions about masculinity and femininity drawn from the Traditional Masculinity Femininity Scale (Kachel, Steffens, & Niedlich, 2016) and two questions about political and religious affiliation drawn from standard Gallup demographic polling.

Procedure

Participants were prompted with a total of 60 sliding scale questions and six free response questions. The sliding scale questions were identical for all participants. Each participant's free response questions were determined by sampling three *personal* questions and three *small talk* questions from the question bank. Questions were displayed in a random order, interleaving free response and sliding scale questions (participants were instructed to expect both formats throughout the task). This trial randomization minimized the risk of free response questions systematically influencing subsequent scale responses or vice versa in a blocked and counterbalanced presentation. Two attention check slider scales identical to those employed in Experiment 1 were inserted at random trial indices (truncating possible trial indices to exclude the first three or last three trials).

In order to achieve roughly similar response lengths for both *small talk* and *personal* free response questions, participants were required to remain on the free response question pages for a minimum of two minutes before they could proceed. A timer at the top of the screen displayed the time remaining and participants were instructed to write as much as possible during the two minutes. There was no time limit enforced on the slider questions. Slider and free response questions were displayed one at a time and once participants had submitted their response, they could not go back.

Results

Participants' free response answers varied considerably in content and tone across questions (Table 1). A linear mixed effects model fit to the character length of responses found

that response lengths differed significantly across question categories ($\chi^2(1) = 6.23$, $p = .01$) with an average difference of around 37 characters (estimated marginal means: *personal*: 270, *se* = 13.7; *small talk*: 233, *se* = 13.7). Participants' responses to the introspection scales also exhibited considerable variance, both between and within scales (see Figure 4 for sample scale responses). Our analyses investigate the degree to which participants' introspection scale responses can be predicted by their free response answers; what do the things they *say* signal about *what they are like*?

To estimate this relationship, we evaluated the accuracy of a large language model—OpenAI's GPT-4o (OpenAI, 2024)—when predicting each participant's slider responses based on their free response answers. To the degree that a language model is able to use information encoded in the free response answers to draw person-specific inferences about personality, this suggests that there is a latent mapping between the content of the free response answers and participants' firsthand evaluations (Park et al., 2024). We tested the language model under three different "conditions" that interrogate the structure of this mapping. First, the language model was asked to predict each participant's slider responses given *all* of their free response answers. Next, the language model was asked to make the same predictions given only the *personal* or the *small talk* answers. Finally, the language model made predictions based on each individual question response alone (we do not analyze these data here). In each of these conditions, the language model was given independent prompts to predict one slider at a time with the relevant evidence. The language model was asked to provide a number between 1 and 100 for the slider; GPT-4o gave an appropriately formatted answer for all requests. We evaluate the accuracy of these predictions under each query condition.

Free response answers carry signal about the speaker When provided with each participant's full set of free response answers, language model predictions for each slider response were modestly correlated with participants' true responses ($r = 0.53$, $p < .0001$; Figure 3B). To quantify this relationship, we fit a linear mixed effects model to the scale predictions with language model estimates as a predictor, while accounting for random variation in participant responses and scale items. Models were fit using *brms* (Bürkner, 2017) and compared according to their *estimated log predictive density* in cross-validation (Vehtari, Gelman, & Gabry, 2017). Our model which includes LLM predictions performs credibly better than a baseline which attributes responses to random variation across participants and scale items (Δelpd $= -980.0$, se $= 49.0$). This suggests that the language model's predictions offer greater predictive accuracy than merely relying on baseline statistical tendencies in people's responses.

Table 1: Sample free response answers.

Depth	Response
Low	"My best pair of shoes are my suede ankle boots. They make me feel like a poet. I feel really myself when I wear them. They are really comfortable, easy to walk in, go with almost everything, and make me feel really cute when I wear them."
High	"I often struggle to admit that I need to slow down or take breaks. I push myself to keep going, fearing that resting might be seen as laziness or a lack of drive, even though it's essential for my well-being."

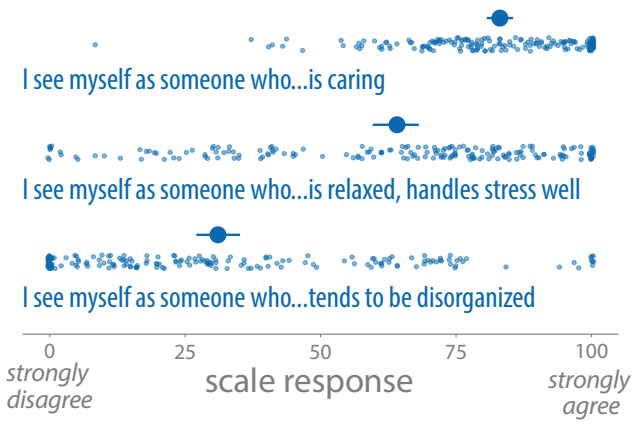

Figure 4: Sample scale responses. Large points show means and 95% bootstrapped confidence intervals. Small points show individual participant judgments.

To further quantify the degree to which the model was relying on participant-specific information encoded in their free response answers, we compared the language model's mean squared prediction error to a null prediction error distribution estimated by repeatedly shuffling scale responses between participants. Any prediction accuracy that can be obtained by merely guessing typical responses to each slider will be preserved in this distribution, while disrupting the relationship between individual participants' free response answers and their corresponding slider values. We find that the observed MSE when the language model is relying on the true mapping between free response answers and slider responses (782.7) falls substantially outside the range of values in our simulated null distribution ($\mu = 1037.6$, $\sigma = 21.4$).

Deep questions support stronger inferences about a speaker Participants' free response answers support prediction of their personality scale responses above and beyond predicting typical values. We next investigated whether this mapping varies for deeper and shallower questions—do answers to deeper questions support stronger inferences about what a speaker is like? To test this hypothesis, we evaluate the language model's prediction accuracy when given *only* participants' responses to the three *personal* questions alongside accuracy given only the *small talk* questions.

First, as above, we estimated a simulated null distribu-

tion for the model under each prediction condition by repeatedly shuffling each participant's scale responses across the existing model predictions. The model's mean squared error for both *personal* (819.6) and *small talk* (877.9) predictions were well outside the range of their respective sampled null distributions (*personal*: $\mu = 1051.7$, $\sigma = 22.1$; *small talk*: $\mu = 1006.9$, $\sigma = 17.5$). This suggests that, overall, information encoded in both the *personal* and *free response* question answers support insights about the respondent above and beyond what might be expected from the model's prior knowledge about people in general.

Next, we investigated whether the language model's predictions differ across these two sources of free response information. We fit a linear mixed effects model to the squared error of model predictions with the *personal* and *small talk* evidence condition as a predictor while accounting for random variation in scale items and participant slider responses in each evidence condition. Our model which included the evidence category (*deep* or *small talk* questions) as a predictor performed credibly better than a baseline which merely attributed squared error to random variation across participants and scale items ($\Delta\text{elpd} = -102.9$, $\text{se} = 17.9$). Estimated marginal mean squared error was lower for *personal* questions than *small talk* questions (*personal*: 815, highest posterior density interval $= [728, 908]$; *small talk*: 874, highest posterior density interval $= [786, 963]$; Figure 3C). This suggests that the language model's ability to draw *person-specific* inferences based on free response answers derives, in part, from differences in the *depth* attributed to the questions.

General Discussion

What kind of questions help us get to know others? We explore the *diagnosticity* of different questions for learning what other people are like. In Experiment 1, we developed a corpus of open-ended questions people could ask in conversation; we investigated participants' judgments about these questions on a variety of scales such as how informative and personal they were. These evaluations of the questions reliably collapsed onto a latent dimension indicating something like the question's interpersonal depth. Agreement among raters about a question's value on this latent dimension provides a first estimate of the question's diagnosticity.

But what information do the answers to these questions contain about a speaker? In Experiment 2, we obtain a second measure of the diagnositicity of questions from Experiment 1 by investigating whether answers to these questions support predictions about the speaker's personality above and beyond what might be true of people in general. We elicited written answers to a subset of questions from Experiment 1, along with a range of standardized sliding scale questions that have been used in prior work on personality, face processing, and demographic profiles (John & Srivastava, 1999; Oosterhof & Todorov, 2008; Kachel et al., 2016). We found that a large language model (OpenAI's GPT-4o) given participants' free response answers was able to predict their sliding scale questions in ways that were person-specific. Furthermore, these predictions were systematically more accurate for questions judged to be high in question depth in Experiment 1, relative to low-depth "small talk" questions. These results provide convergent evidence for the questions' interpersonal depth and suggest that intuitive notions of a question's depth may reflect latent, abstract information about a speaker that can be inferred from the their answer to the question.

Nonetheless, the difference in predictive accuracy across deep and small talk questions was relatively small compared to the difference in *perceived* depth of those same questions. This may have resulted from the fact that participants were forced to spend the same amount of time answering all the questions, in effect leading to "deeper" answers to the small talk questions (Table 1); future work eliciting evaluations of the free response *answers* similar to those provided for the questions in Experiment 1 offers the potential to further quantify the relationship between question and answer depth.

In addition, the current results offer a number of avenues for clarifying how people learn what others are like using natural language. The ability of a language model to draw person-specific inferences about a respondent indicates that information about a speaker is present in their answers to particular questions. How accurately can people extract this information from the same answers? In ongoing work, we evaluate human participants' ability to predict aspects of a speaker's personality from answers to deep and small talk questions. Human predictions of the sort made by GPT-4o in the current results may offer insights into the social inferences involved in getting to know others.

Insofar as people are able to draw insights about a speaker from natural language, the question remains *how* we do this. What are sorts of abstractions we use to represent what others are like and how do we acquire this information when talking with people? Recent work exploring these questions (van Baar, Nassar, Deng, & FeldmanHall, 2022; FeldmanHall & Shenhav, 2019; FeldmanHall & Nassar, 2021; Tamir & Thornton, 2018) may benefit from the approach in the current results; in particular, a more fine-grained understanding of the inferences people make about others from their answers to open-ended questions may enable concrete hypotheses about people's causal models of how personality impacts behavior; our *intuitive personality psychology*.

Further, understanding the inferential processes at play when interpreting others' answers to deep and small talk questions may allow for novel theories about the dynamics of human conversation. What role do epistemic goals of trying to learn about others play in our interactions with them and how do people engage in this sort of *active learning about others* over the course of everyday interactions? More broadly, the approach taken here may point toward a better understanding of why humans spend so much time and energy *getting to know one another*, and how we use language to achieve this.

Acknowledgments

E.B. is supported by NSF SBE Postdoctoral Research Fellowship #2404706. T.G. is supported by grants from Stanford's Human-Centered Artificial Intelligence Institute (HAI) and Cooperative AI. J.E.F. is supported by NSF DRL award #2400471 and NSF CAREER award #2047191 and an ONR Science of Autonomy award.

References

Allport, G. W. (1961). *Pattern and growth in personality.* Holt, Reinhart & Winston.

Aron, A., Melinat, E., Aron, E. N., Vallone, R. D., & Bator, R. J. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and social psychology bulletin*, *23*(4), 363–377.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Cervone, D., & Pervin, L. A. (2022). *Personality: Theory and research*. John Wiley & Sons.

Costa Mastrascusa, R., de Oliveira Fenili Antunes, M. L., de Albuquerque, N. S., Virissimo, S. L., Foletto Moura, M., Vieira Marques Motta, B., ... Quarti Irigaray, T. (2023). Evaluating the complete (44-item), short (20-item) and ultra-short (10-item) versions of the big five inventory (bfi) in the brazilian population. *Scientific Reports*, *13*(1), 7372.

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature communications*, *10*(1), 2319.

FeldmanHall, O., & Nassar, M. R. (2021). The computational challenge of social learning. *Trends in Cognitive Sciences*, *25*(12), 1045–1057.

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature human behaviour*, *3*(5), 426–435.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American psychologist*, *48*(1), 26.

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, *37*(6), 504–528.

Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the big five? testing the five-factor model of personality variation among forager–farmers in the bolivian amazon. *Journal of personality and social psychology*, *104*(2), 354.

John, O. P., & Srivastava, S. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), (2nd ed., pp. 102–138). Guilford Press.

Kachel, S., Steffens, M. C., & Niedlich, C. (2016). Traditional masculinity and femininity: Validation of a new scale assessing gender roles. *Frontiers in psychology*, *7*, 956.

Mahaphanit, W., Welker, C., Schmidt, H., Chang, L., & Hawkins, R. (2024). When and why does shared reality generalize? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

OpenAI. (2024). *OpenAI o1 System Card* (Tech. Rep.). Retrieved 2024-10-16, from https://cdn.openai.com/o1-system-card-20240917.pdf

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., ... Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, *41*(1), 203–212.

Rothbart, M. K., Ahadi, S. A., & Evans, D. E. (2000). Temperament and personality: origins and outcomes. *Journal of personality and social psychology*, *78*(1), 122.

Shiner, R. L., Soto, C. J., & De Fruyt, F. (2021). Personality assessment of children and adolescents. *Annual Review of Developmental Psychology*, *3*(1), 113–137.

Steyn, R., & Ndofirepi, T. M. (2022). Structural validity and measurement invariance of the short version of the big five inventory (bfi-10) in selected countries. *Cogent Psychology*, *9*(1), 2095035.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in cognitive sciences*, *22*(3), 201–212.

van Baar, J. M., Nassar, M. R., Deng, W., & FeldmanHall, O. (2022). Latent motives guide structure learning during adaptive social choice. *Nature Human Behaviour*, *6*(3), 404–414.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.