Elicitation Inference Optimization for Multi-Principal-Agent Alignment

Anonymous Author(s)
Affiliation
Address
email

Abstract

In multi-principal agent alignment scenarios spanning governance, markets, diplomacy, and AGI, it is unfeasible to elicit every principal's view on all perspectives relevant to agent decisions. Elicitation inference optimization (EIO) aims to minimize the n elicitations needed to approximate N principal's views across K perspectives. In this work, we demonstrate an EIO approach where data efficiency (NK/n) increases with scale. We introduce STUMP: an elicitation inference model which integrates an LLM with a latent factor model to enable learning transfer across samples, contexts, and languages. Then, we characterize STUMP's performance on a set of elicitation primitives from which scalable elicitation (sampling) protocols can be constructed. Building from these results, we design and demonstrate two scalable elicitation protocols for STUMP where data efficiency grows boundlessly, scaling like $O(n)$ in the number of elicitations n. This makes it possible to obtain complex, high-dimensional preference signals spanning principal populations at any scale.

1 Introduction

The principal agent problem involves aligning agent decisions with principal interests. A challenge is creating situations where agent choices are sufficiently influenced by signals containing principal preferences. With a single principal, high-complexity preference signals can be elicited directly via open-ended interaction. Multi-principal agent scenarios can involve large populations of principals and powerful agents such as: governments & citizens[1, 2], firms & customers[3], peacekeepers & conflict parties[4], and potentially even AGI & humanity [5, 6]. As the number of principals grows, and the domain of agent decisions becomes open-ended, directly eliciting the preference of all principals on all relevant perspectives becomes unfeasible [A.1]. As a result, lower-complexity forms of elicitation like ballot voting (ie. for governments) and price signals (ie. for firms) are used to learn preferences. While clearly effective—a basis of democracy and the economy—these approaches drastically simplify real preferences. For example, they do not allow citizens (principals) to express *what* they would like a government (agent) to do, or *why*, only *if* they support predefined options.

Elicitation inference optimization (EIO) aims to decrease the amount of direct elicitation needed to recover a preference signal (enabling use of more complex, higher-dimension preferences; e.g. in natural language). Consider an $N \times K$ matrix Θ where rows corresponds to N principals, columns corresponds K perspectives, and every element captures a principal-perspective relationship. *The goal of elicitation inference optimization* is to obtain a sufficient approximation of Θ with a minimal elicitation budget by directly sampling some elements and inferring the rest. Thus, EIO involves combining a) a sparse elicitation (sampling) protocol with, b) an elicitation inference model.

Closed-ended surveys simplify EIO by constraining the set of relevant perspectives to a predefined set — typically, with $K << N$. Matrix sampling techniques elicit responses from each

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

participant on a subset of perspectives selected randomly [7], heuristically [8], or dynamically such that inference accuracy is adaptively optimized [9, 10]. Inference models exploit the low-rank-ness of Θ via matrix factorization and other collaborative filtering techniques [10, 11, 12].

Collective dialogue systems (CDS) enable arbitrarily high-complexity signals by allowing the set of relevant perspectives to be open-ended: neither limited in number nor requiring pre-definition. In a CDS, participants contribute open-ended perspectives in the context of a question, prompt, or conversation and respond to subsets of perspectives contributed by others [13] – typically, yielding $K \sim O(N)$. Inspired by preference models [14] and conjoint analysis [15], one class of approaches elicits randomized pair-choice votes [16] and uses hierarchical probit-like models for inference [17]. Other approaches elicit agreement votes to learn absolute human-perspective relationships. Pol.is samples agreement votes semi-randomly, prioritizes votes that improve clustering, and uses mean imputation to support dimensionality reduction [18]. Additional work bridges these approaches by randomly sampling both pair-choice and agreement votes; inferring missing votes by modeling both votes types with a single utility matrix learned via regularized matrix completion [19].

Previous approaches to EIO for CDS have been limited in scope to single samples where the set of relevant perspectives all correspond to the same context (ie. the same question or prompt) [17, 16, 18, 19]. This means votes must be elicited from every person for every new context. As a result, to approximate some Θ spanning an arbitrary number of contexts, the number of direct elicitation's needed grows proportional to the number of elements in Θ [A.2] – data efficiency does not increase with scale. However, previous approaches do not fully leverage all available data: learning from one context is not transferred to support inference in other contexts, and the information contained in perspective text is not used it all. *Can an EIO approach which better leverages available data enable increasing data efficiency with scale?*

In this work we introduce an approach to EIO for CDS which becomes increasingly data efficient as the amount of data elicited grows. First, we introduce a novel elicitation inference model – STUMP – which better leverages available data by integrating a pre-trained LLM with a latent factor model. Second, we characterize STUMP's performance across a range of elicitation primitives from which arbitrarily scalable elicitation (vote sampling) protocols can constructed. Finally, building on these results, we design and demonstrate two arbitrarily scalable elicitation protocols where STUMP can infer with meaningful accuracy while data efficiency increases linearly with scale.

2 Problem Setup

Elicitation form. In a collective dialogue system (CDS) participants respond to prompts with open-ended perspectives and vote on perspectives submitted by others. Let H be the set of N human participants, P be the set of K perspectives, and Θ be the corresponding $N \times K$ human-perspective matrix. Both agreement and pair-choice votes may be elicited on perspectives. In agreement exercise e^a_{ij}, participant i is asked if they agree with perspective j. In pair choice exercise e^c_{ijk}, participant i is asked if they prefer perspective j or k. We denote a set of exercise data $E = \{e^a_{ij}, e^c_{ijk} \mid i \in H; j, k \in P\}$.

Elicitation inference involves training a model on available data and predicting all missing elements in a target Θ. We choose binary agreement as the human-perspective relationship for our target Θ. Inference performance is probed experimentally by training on a subset of exercises including both vote types E_t, and computing prediction accuracy on a validation set of only agreement votes E_v.

Data efficiency in EIO can be quantified as the ratio of a) the number of votes contained in the target Θ and b) the number of directly elicited votes a model needs trained on to meaningfully approximate Θ. We denote the *data efficiency* to achieve inference accuracy acc as:

$$\beta_{acc} = \frac{n(\Theta)}{n(E_t)} = \frac{n(H) \times n(P)}{n(E_t)} = \frac{NK}{n_t}$$

where $n(X)$ denotes the size of set X and $n_t \equiv n(E_t)$. We refer to the rate data efficiency changes with more elicited votes ($\partial\beta_{acc}/\partial n_t$) as *data leverage*. EIO becomes more data efficient with scale if data leverage is positive for arbitrarily large n_t above a sufficient threshold of accuracy. To achieve this, the challenge is to combine: a) an arbitrarily scalable elicitation (sampling) protocol where

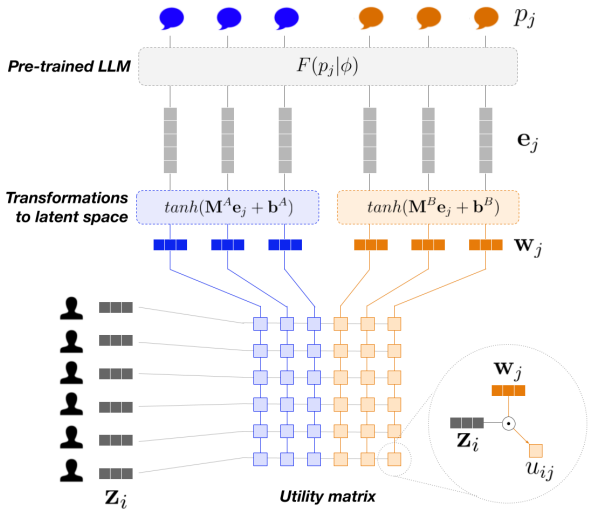

Figure 1: Diagram of STUMP. Universal model elements shown in grey. Perspectives and context-specific model elements colored by context.

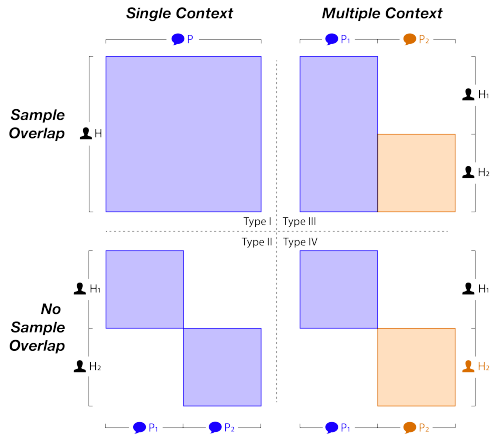

Figure 2: Elicitation primitives. Color boxes indicate sparse vote data available for training. Colors denote context.

NK grows faster than n_t, and b) an inference model that sustains meaningful inference accuracy at arbitrary scale using the data the protocol generates.

3 Elicitation Inference Model

Elicitation inference (EI) is analogous to collaborative filtering (CF) in recommenders, except instead of items people may prefer, there are *perspectives* people may agree with. A standard approach to CF is to learn a low-rank approximation of a user-item preference matrix by optimizing an objective function against observed preferences [20]. Bilich et al [19] demonstrated a similar approach for EI on CDS data. They use regularized matrix completion with a low-rank-promoting nuclear norm prior (similar to [21]) to learn a utility matrix which predicts votes. We refer to this as the *nuclear norm* model (NN), and adopt it as a baseline (see A.3 for technical details).

Latent factor models (LFMs) learn vectors for each user $\mathbf{z}_i$ and item $\mathbf{w}_j$ which embed them in the same latent space, and describe the preference of user i for item j as a function of $u_{ij} = \mathbf{z}_i \cdot \mathbf{w}_j$ [20]. Known item features (i.e., director of a movie) can be used to improve performance by projecting features into latent space as a component of item embeddings [22]. Relevant features can also be extracted from item content (i.e., a song's time sequence) by a link to latent space embeddings [23].

We consider EI on text-based perspectives. Collaborative Topic Ranking (CTR) uses information in text by learning a topic model on item texts and equating topic feature vectors to latent space embeddings [24]. Pre-trainable large language models (LLMs) [25, 26] learn to encode information from text into embeddings suitable for a wide range of down stream tasks [27, 28, 29]. Our approach to use information in perspective texts is to start with a pre-trained LLM, then similar to [22], learn a transformation of text embeddings into latent space. A key challenge in CF is maintaining alignment between latent spaces learned from non-overlapping data [30]. Sharing the same embedding model and transformation across perspectives enforces alignment even in the absence of overlapping data (see section 4.3).

Cross-Domain recommendation (CDR) involves aligning latent spaces across items in different domains like movies and songs [31]. One class of approaches use information like tags that can be shared across domains [32]. Others focus on aligning latent space distributional properties [33].

In EI the analog to CDR is multi-context inference. How one votes on a perspective depends on the context it was elicited in. Take the perspective "*I feel happy.*" Consider two scenarios, where it was elicited by asking "*What would you say if you saw your parents smile?*" or "*What wouldn't you say if you saw your parents smile?*". One would expect near opposite voting behavior. But, if

the same embedding model and transformation are shared in both contexts, then predictions will be the same. Enabling multi-context inference requires relaxing this condition. In the spirit of [32] our approach is to share the same LLM across contexts, but learn a unique transformation into latent space for each context.

Our model. We formally introduce our model, the Semantic Transfer Utility Model of Participants: STUMP. We denote the event that participant i reports agreement with perspective j by a_{ij}, and disagreement by d_{ij}. We denote the event that participant i reports preference for perspective j over perspective k by c_{ijk}. We assume a logit utility model of choice (similar to [20, 19]) where participant i has a latent utility u_{ij} corresponding to perspective j and voting probabilities are given by $p(a_{ij}|u_{ij}) = \sigma(u_{ij})$, $p(d_{ij}|u_{ij}) = \sigma(1 - u_{ij})$ and $p(c_{ijk}|u_{ij}) = \sigma(u_{ij} - u_{ik})$, where σ denotes the logistic function. The utility u_{ij} for participant i and perspective j is constructed as an LFM:

$$u_{ij} = \mathbf{z}_i \cdot \mathbf{w}_j$$

where $\mathbf{z}_i$ is a latent embedding for participant i, and $\mathbf{w}_j$ is a latent embedding for perspective j. To use information in perspective text, we employ a pre-trained LLM [34, 25] to generate text embeddings. We denote the embedding of the jth perspective text p_j by model F with parameters ϕ, as:

$$\mathbf{e}_j = F(p_j|\phi).$$

Text embeddings are mapped into latent space by context-specific transformations:

$$\mathbf{w}_j = tanh(\mathbf{M}^q \mathbf{e}_j + \mathbf{b}^q)$$

where $\mathbf{M}^q$ and $\mathbf{b}^q$ are the transformation weight and bias for context q (see Figure 1). Finally, let $A = \{i, j|a_{ij}\}$, $D = \{i, j|d_{ij}\}$, and $C = \{i, j, k|c_{ijk}\}$. The log loss is defined by:

$$L = -(\sum_{i,j,k \in C} log(\sigma(u_{ij} - u_{ik})) + \sum_{i,j \in A} log(\sigma(u_{ij})) + \sum_{i,j \in D} log(1 - \sigma(u_{ij}))) + \lambda_m \|\mathbf{M}\|_F$$

where $\|.\|_F$ denotes the Frobenius norm of a matrix, and λ_m is a hyper-parameter controlling regularization strength. The participant embedding vectors $\mathbf{z}$ and the set of transformation parameters for each context $\{\mathbf{M}^q, \mathbf{b}^q\}$ are learned by minimizing the log-loss for observed data. Optionally, the LLM parameters ϕ may be fine-tuned through the same minimization.

4 Elicitation Primitives

Our strategy for elicitation protocol design begins with characterizing STUMP's performance across a range of elicitation primitives from which arbitrarily scalable protocols can be constructed. To this end, we explore four types of elicitation primitives arising from two orthogonal data set properties: *context homogeneity* and *sample overlap*.

Context homogeneity. Perspectives and votes are elicited under some context — for example, in response to a question. Thus, a data set (or θ) may correspond to one or more context. We denote the set of perspectives corresponding to the qth context as $P^q \subseteq P$.

Sample overlap. Vote data E may come from one or more samples, where each sth sample involves a subset of humans $H_s \subseteq H$ and a subset of perspectives $P_s \subseteq P$. Two samples are said to have *no overlap* when they share neither humans nor perspectives, ie $H_1 \cap H_2 = \emptyset$ and $P_1 \cap P_2 = \emptyset$. A sample with *overlap* may correspond to a data set with only a single sample, or with multiple samples where either humans or perspectives are shared, ie $H_1 \cap H_2 \neq \emptyset$ or $P_1 \cap P_2 \neq \emptyset$.

Permuting these two properties gives four types of elicitation primitives (Figure 2):

Type	Context	Sample	Definition
I	Single	Overlap	$P^q = P$ and $H_1 \cap H_2 \neq \emptyset$ or $P_1 \cap P_2 \neq \emptyset$
II	Single	No overlap	$P^q = P$ and $H_1 \cap H_2 = P_1 \cap P_2 = \emptyset$
III	Multiple	Overlap	$P^q \neq P$ and $H_1 \cap H_2 \neq \emptyset$ or $P_1 \cap P_2 \neq \emptyset$
IV	Multiple	No overlap	$P^q \neq P$ and $H_1 \cap H_2 = P_1 \cap P_2 = \emptyset$

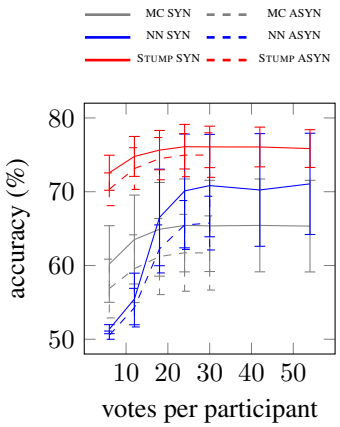

Figure 3: Type I inference accuracy vs votes per participant.

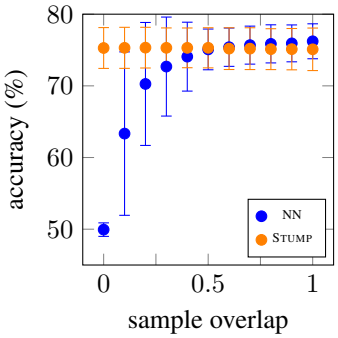

Figure 4: Type I/II inference accuracy vs sample overlap.

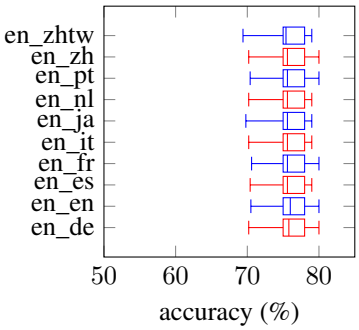

Figure 5: Type II inference accuracy by language pair.

4.1 Elicitation primitive experiments

Baselines. We use three prediction baselines to contextualize STUMP's performance: a) a *Random* baseline where votes are predicted randomly with uniform probability, b) a *Most Common* (MC) baseline where votes are predicted to be the most common choice observed for each perspective, and c) the *nuclear norm* (NN) model from Bilich et al [19] [A.3] which represents a standard approach to CF and is the only prior work we are aware of focused on single-vote-level inference for CDS.

Data and implementation. Data used for experiments was collected as part of regular load testing for an existing CDS and de-identified prior to use in this study. Data came from 534 participants solicited through online research panels. We use data corresponding to seven questions [A.9] asked during the CD , chosen for their nuance and presumed hard-to-predict nature. Each question is treated as a unique context, meaning only perspectives and votes elicited in response to the same question share the same context transformation weights. Experiments are constructed by partitioning subsets of data into training and validation sets (E_t and E_v). Data and training details are described in [A.4].

4.2 Type I: Single context, overlapping sample

Data elicited from one question during a collective dialogue (CD) represents a type I primitive. We test type I model performance as a function of *vote per participant* (VPP) for two kinds of CD; asynchronous and synchronous. In *asynchronous* CDs, participation happens on a rolling basis, so participants only vote on perspectives submitted prior to their participation. In *synchronous* CDs, all participation happens in the same minute-scale window of time, so any perspective may be voted on by any participant. We simulate both cases while varying VPP by partitioning the appropriate votes into training and validation sets [A.5.1]. For each experiment we train the models for each question and compute the corresponding validation accuracy. STUMP outperforms all baselines at all VPP for both cases with < 10 VPP and reaches peak performance at ~ 20 VPP (Figure 3).

Type I as a baseline. As the simplest elicitation primitive, Type I performance across baselines informs what constitutes *meaningful accuracy* in general. The *Random* baseline achieves 50% accuracy (not shown), the *Most Common* baseline saturates at 64% accuracy, and the *NN* baseline at 70% accuracy. A reasonable conclusion is that $50 - 64\%$ accuracy is not meaningful, $65 - 69\%$ is arguably meaningful, and 70% or more is meaningful.

4.3 Type II: Single context, non-overlapping sample

Type I to Type II transition. We simulate EI within a single-question context as the overlap between two samples decreases to zero. For each question, we randomly split the participants and perspectives into two equal sized samples: $\{H_{s1}, P_{s1}\}$ and $\{H_{s2}, P_{s2}\}$. Agreement votes by participants in H_{s1} on perspectives in P_{s2} become the validation set. The remaining votes become the first training

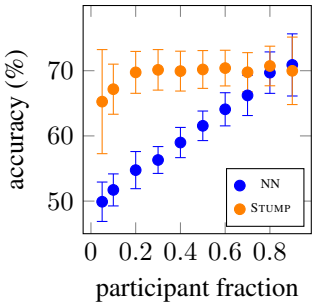

Figure 6: Type III inference accuracy vs participant fraction r.

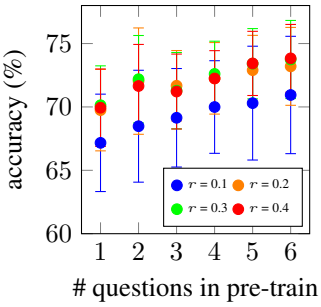

Figure 7: Type III inference accuracy vs questions in pre-training data.

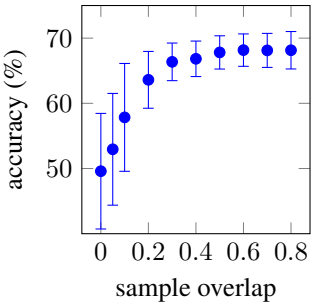

Figure 8: Type III/IV inference accuracy vs sample overlap.

set, where sample overlap is maximum (Type I). Sample overlap is decreased by removing votes on perspectives in P_{s1} by participants in H_{s2} until none remain (Type II) [A.5.2]. We define sample overlap p as the fraction of participants in H_{s2} with votes in P_{s1}, and compute validation accuracy for STUMP and the NN baseline as p varies (Figure 4). In contrast to the NN baseline, STUMP's performance remains stable even at $p = 0$; *evidence of* STUMP*'s ability to maintain latent-space alignment across non-overlapping samples.*

Type II with Different languages. We investigate STUMP's inference capability across two non-overlapping samples in different languages. Perspectives in one sample are translated into one of nine different languages using Google Translate API: German (de), Spanish (es), French (fr), Italian (it), Japanese (ja), Dutch (nl), Portuguese (pt), simplified Chinese (zh) and traditional Chinese (zhtw) [A.5.2]. We compute STUMP's Type II accuracy for each language pair and compare it to an English-English baseline (Figure 5). Performance across all language pairs is on par with the baseline.

4.4 Type III: Multiple context, overlapping sample

Transfer learning. To test STUMP's capability for transfer learning, we pre-train participant embeddings in one context, and use them for inference in a new context. The pre-training set contains votes by all participants on perspectives in context Q_1 and is used to pre-train participant embeddings. The second training set contains votes for a fraction r of participants on perspectives in a new context Q_2. The pre-trained embeddings are frozen while the transformation to latent space for the new context is learned. Votes on perspectives in Q_2 by participants not in the second training set comprise the type III validation set. We vary the *participant fraction* r from 0.05 to 0.9, and compute validation accuracy for STUMP and the NN baseline (Figure 6). STUMP performance increases with participant fraction when it is small, leveling off at $r \sim 0.2$, while the NN baseline only reaches similar performance around $r \sim 0.8$. In other words, STUMP is able to leverage data from a different context via transfer learning to achieve the same performance as the NN baseline with 75% less data.

Increased pre-training data. Next, we investigate how more pre-training data from more contexts effects STUMP's transfer learning performance. We increase the number of question contexts whose votes are included in the pre-training set [A.5.3] and compute type III validation accuracy at different participant fractions (Figure 7). STUMP's transfer learning performance increases as the number of question contexts included at pre-training increases.

4.5 Type IV: Multiple context, non-overlapping sample

Type III to Type IV transition. Finally, we explore transfer learning across contexts as sample overlap decreases. Let *sample overlap* p be the fraction of participants in the training sample (context Q_2) who are also in the pre-training sample (context Q_1). Moving from Type III to Type IV means decreasing p from 1 to 0. To simulate this, we start with a Type III setup and iteratively remove participants from the pre-training sample who are also in the training sample [A.5.4]. We use the training procedure and validation set from [4.4] and compute accuracy as p varies (Figure 8). STUMP performance is arguably meaningful when sample overlap is larger than 50% but drops below that.

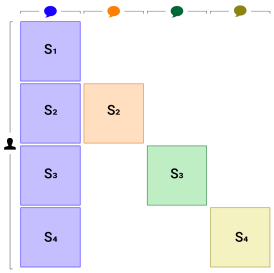

Figure 9: Case 1 sampling protocol, where shaded regions correspond to sampled votes and color corresponds to context.

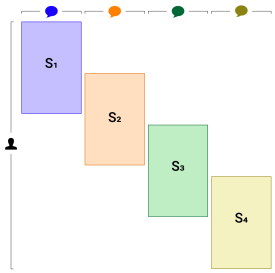

Figure 10: Case 2 sampling protocol, where shaded regions correspond to sampled votes and color corresponds to context.

5 Increasing Data Efficiency with Scale

Based on STUMP's performance characteristics across elicitation primitives, we design two arbitrarily scalable elicitation protocols where: a) one may expect STUMP to sustain meaningful inference accuracy, and b) data leverage is positive, ie. $\frac{\partial \beta}{\partial n_t} > 0$.

Case I. STUMP shows meaningful type III inference for low participant fractions (Figure 6). This means if participants all vote on perspectives from a shared context, only a small fraction need to vote on perspectives from a new context to meaningfully infer the rest. Thus, the following elicitation protocol is viable: Elicit an initial sample where m participants each give 1 new perspective and vote on c other perspectives for a single context. Let each sample after the first be elicited from m new participants where each a) provides 1 new perspective and votes on c other perspectives for a new context, and b) votes on c perspectives for the context shared by all samples (Figure 9). This protocol results in data efficiency of $\beta = n_t/(4c^2) + m/(2c) + m/(4n_t) \sim O(n_t)$, and data leverage $\frac{\partial \beta}{\partial n_t} \approx 1/(4c^2) > 0$. [A.6.1]

Case II. STUMP shows arguably meaningful type III/IV inference once sample overlap approaches ~ 0.5 (Figure 8). In other words, inference for a population of participants across two contexts requires around half of them to vote on perspectives from both contexts. This implies such an overlap supports latent space alignment between contexts. To maintain latent space alignment across an arbitrary number of contexts, we chain successive overlaps together with the following elicitation protocol: Elicit an initial sample where m participants each provide 1 new perspective and vote on c other perspectives for a single context. Let each sample after the first be elicited from m participants where $m/2$ participants are new, and $m/2$ are from the previous sample. Let each participant provide 1 new perspective and vote on c other perspectives for a new context (Figure 10). This protocol results in data efficiency of $\beta = n_t/(2c^2) + m/(2c) \sim O(n_t)$, and $\frac{\partial \beta}{\partial n_t} = 1/(2c^2) > 0$. [A.6.2]

We simulate EI under both protocols for $q = 2, 3, 4$ samples by partitioning the appropriate perspectives, participants, and votes into training and validation sets [A.7]. In both cases STUMP sustains a meaningful accuracy of $\sim 70\%$ as q increases (Figure 11), approximating $O(n_t^2)$ votes using only $O(n_t)$. Thus, the combination of either elicitation protocol with STUMP achieves $\beta_{70} \sim O(n_t)$ (Figure 12), giving positive data leverage: $\frac{\partial \beta_{70}}{\partial n_t} = const > 0$. In other words, *data efficiency increases linearly with scale.*

6 Discussion

Improving inference accuracy. This work prioritized data efficiency with scale over maximizing inference accuracy. Experiments in (Section 4.4) indicate more, and more diverse, pre-training data can improve accuracy. STUMP's use of information in perspective text is limited by the LLM it employs and we expect a SOTA LLM [35, 36, 37] can improve accuracy. Other directions to improve accuracy may include: using participant side information like demographics, extending STUMP to use information in prompts, using active sampling to elicit votes expected to give the most information [9, 10], and fine-tuning LLM weights while training STUMP.

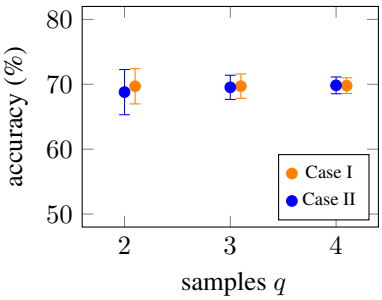

Figure 11: Inference accuracy as a function of the number of local samples.

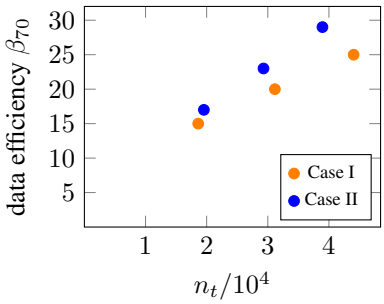

Figure 12: Data leverage as a function of the number of votes in the training set.

Limitations. STUMP fails to align latent spaces across non-overlapping samples from different contexts. This may require an embedding distribution alignment module and Stien Path training [33]. Our work does not quantify confidence in aggregate vote statistics, though [19] shows stochastic weight averaging [38] can estimate posterior variance in this setting. We only use text perspectives. Handling perspectives spanning text, images, audio, and video requires swapping STUMP's LLM for a multi-modal foundation model able to generate aligned embeddings across media types [39, 40, 41, 42, 43]. STUMP can only represent changing participant views by treating different times as different contexts, while recommender often model it explicitly [44]. This work does not show how to make results understandable for human agents, though clustering in latent space is a promising direction [18]. We don't show how learned preference signals can directly control other ML models, but jury learning for classifiers [45] and reward learning for RL agents [46] are encouraging directions. STUMP is not deferentially private [47] but approaches developed for LFMs in recommenders may be applicable [48]. Finally, elicitation inference is just one of the components helpful for achieving the underlying goals of supporting alignment.

Risks. In CDS, some biases common to recommenders [49] are less significant due to the structure of the task. In particular, if vote exercises are random and can't be skipped, while users are anonymous and can't see results, then biases spanning selection, exposure, conformity, position, and popularity are minimized. However, participants who respond and vote more slowly may still be under-represented in the data, and bad elicitation prompts can still bias participants [50]. Biases in pre-trained LLMs [51, 52, 53] may also propagate through the model to effect inference in unknown ways. We avoid active sampling [10] due to the risk of compounding small model bias into large inference bias. Further, non-representative participant populations can arise from biased solicitation, or unequal willingness or ability to participate. As such systems become used for important decision-making, they risk neglecting views of already marginalized communities due to limited technology access and capacity [54] unless specifically mitigated for, e.g. using traditional voting infrastructure, increasing digital equity, and/or sortition with high touch support.

Conclusion. We have demonstrated EIO for CDS where data efficiency increases linearly with the size of the elicited data set; our approach approximates $O(n^2)$ votes while requiring only $O(n)$ be elicited directly. Our experiments suggest this is likely to hold for arbitrarily large data sets spanning contexts, languages, and population samples. This implies data efficiency may increase boundlessly with scale, making it feasible to approximate high-complexity signals for massive principal populations spanning open-ended perspectives across an agent's decision domain. These signals may be used to steer complex objective and evaluation functions for multi-principal agent scenarios spanning governance, markets, diplomacy, and AGI alignment. More speculatively, as foundation models (FMs) increasingly demonstrate capabilities spanning those of LLMs and RL agents [41, 43, 55, 56, 57], EIO may help enable a scenario where the quality and complexity of signals driving an AGI's alignment increases as its capability grows A.8.

References

[1] Nathalie Giger and Zoe Lefkofridi. Salience-based congruence between parties & their voters: The swiss case. *Swiss political science review*, 20(2):287–304, 2014.

[2] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437, 2020.

[3] Nicholas Roberts and Varun Grover. Investigating firm's customer agility and firm performance: The importance of aligning sense and respond capabilities. *Journal of Business Research*, 65(5): 579–585, 2012.

[4] Strengthening the role of mediation in the peaceful settlement of disputes, conflict prevention and resolution. URL `https://peacemaker.un.org/sites/peacemaker.un.org/files/SGReport_StrenghteningtheRoleofMediation_A66811_0.pdf`.

[5] Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4):105–114, 2015.

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf`.

[7] David M Shoemaker. *Principles and procedures of multiple matrix sampling.* Ballinger, 1973.

[8] Trivellore E Raghunathan and James E Grizzle. A split questionnaire survey design. *Journal of the American Statistical Association*, 90(429):54–63, 1995.

[9] Jeffrey M Gonzalez and John L Eltinge. Adaptive matrix sampling for the consumer expenditure quarterly interview survey. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 2081–2088, 2008.

[10] Chelsea Zhang, Sean J Taylor, Curtiss Cobb, and Jasjeet Sekhon. Active matrix factorization for surveys. *The Annals of Applied Statistics*, 14(3):1182–1206, 2020.

[11] Nandana Sengupta, Nati Srebro, and James Evans. Simple surveys: Response retrieval inspired by recommendation systems. *Social Science Computer Review*, 39(1):105–129, 2021.

[12] Markos FBG Oliveira, Myriam Delgado, and Ricardo Lüders. Comparative analysis of collaborative filtering-based predictors of scores in surveys of a large company. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 739–750. SBC, 2021.

[13] Collective dialogue systems. URL `https://www.collectivedialogue.com/`.

[14] Louis L Thurstone. The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21(4):384, 1927.

[15] Paul E Green and Venkatachary Srinivasan. Conjoint analysis in consumer research: issues and outlook. *Journal of consumer research*, 5(2):103–123, 1978.

[16] Andrew Konya and Aaron Slodov. Collective speech.

[17] Matthew J Salganik and Karen EC Levy. Wiki surveys: Open and quantifiable social data collection. *PloS one*, 10(5):e0123483, 2015.

[18] Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. Polis:: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca. Revista de Pensament i Anàlisi*, 26(2):1–26, 2021.

[19] Jordan Bilich, Michael Varga, Daanish Masood, and Andrew Konya. Faster peace via inclusivity: An efficient paradigm to understand populations in conflict zones.

[20] Christopher C. Johnson. Logistic matrix factorization for implicit feedback data. 2014.

[21] M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference*, volume 6.

[22] Haiyang Zhang, Ivan Ganchev, Nikola S. Nikolov, Zhanlin Ji, and Máirtín O'Droma. Featuremf: An item feature enriched matrix factorization model for item recommendation. *IEEE Access*, 9: 65266–65276, 2021. doi: 10.1109/ACCESS.2021.3074365.

[23] Aaron van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL `https://proceedings.neurips.cc/paper/2013/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf`.

[24] Xin Xin, Zhirun Liu, Chin-Yew Lin, Heyan Huang, Xiaochi Wei, and Ping Guo. Cross-domain collaborative filtering with review text. 07 2015.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[26] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Multilingual universal sentence encoder for semantic retrieval. *CoRR*, abs/1907.04307, 2019. URL `http://arxiv.org/abs/1907.04307`.

[27]

[28] Boran Hao, Henghui Zhu, and Ioannis Paschalidis. Enhancing clinical BERT embedding using a biomedical knowledge base. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 657–661, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.57. URL `https://aclanthology.org/2020.coling-main.57`.

[29] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *EACL*, 2021.

[30] Paolo Cremonesi and Massimo Quadrana. Cross-domain recommendations without overlapping data: Myth or reality? In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 297–300, 2014.

[31] Tianzi Zang, Yanmin Zhu, Haobing Liu, Ruohan Zhang, and Jiadi Yu. A survey on cross-domain recommendation: Taxonomies, methods, and future directions. *ACM Transactions on Information Systems*, 07 2022. doi: 10.1145/3548455.

[32] Ashish K Sahu, Pragya Dwivedi, and Vibhor Kant. Tags and item features as a bridge for cross-domain recommender systems. *Procedia Computer Science*, 125:624–631, 2018.

[33] Weiming Liu, Jiajie Su, Chaochao Chen, and Xiaolin Zheng. Leveraging distribution alignment via stein path for cross-domain cold-start recommendation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19223–19234. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/a0443c8c8c3372d662e9173c18faaa2c-Paper.pdf`.

[34] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[35] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani

Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022.

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

[37] Bloom. URL `https://huggingface.co/bigscience/bloom`.

[38] Wesley J. Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. *A Simple Baseline for Bayesian Uncertainty in Deep Learning*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[39] Muhammad Shahid Jabbar, Jitae Shin, and Jun-Dong Cho. Ai ekphrasis: Multi-modal learning with foundation models for fine-grained poetry retrieval. *Electronics*, 11(8), 2022. ISSN 2079-9292. doi: 10.3390/electronics11081275. URL `https://www.mdpi.com/2079-9292/11/8/1275`.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

[41] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2021. URL `https://arxiv.org/abs/2108.07258`.

[42] Ruihua Song, Xin Gao, Tao Xiang, Hao Sun, and Ji-Rong Wen. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(13), 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30761-2. URL `https://doi.org/10.1038/s41467-022-30761-2`.

[43] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent, 2022. URL `https://arxiv.org/abs/2205.06175`.

[44] Yi-Cheng Chen, Lin Hui, and Tipajin Thaipisutikul. A collaborative filtering recommendation system with dynamic time decay. *The Journal of Supercomputing*, 77, January 2021. doi: 10.1007/s11227-020-03266-2.

[45] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeffrey T. Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. *CoRR*, abs/2202.02950, 2022. URL https://arxiv.org/abs/2202.02950.

[46] Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593, 2019.

[47] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[48] Arnaud Berlioz, Arik Friedman, Mohamed Ali Kaafar, Roksana Boreli, and Shlomo Berkovsky. Applying differential privacy to matrix factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 107–114, 2015.

[49] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ArXiv*, abs/2010.03240, 2020.

[50] Eugene Litwak. A classification of biased questions. *American Journal of Sociology*, 62(2): 182–186, 1956. doi: 10.1086/221960. URL https://doi.org/10.1086/221960.

[51] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4:258–268, 03 2022. doi: 10.1038/s42256-022-00458-8.

[52] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

[53] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654, 2022. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103654. URL https://www.sciencedirect.com/science/article/pii/S0004370221002058.

[54] Philip Roessler. The mobile phone revolution and digital inequality: Scope, determinants and consequences. *Prosperity Commission Background Paper Series*, (15), 2018.

[55] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15084–15097. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/7f489f642a0ddb10272b5c31057f0663-Paper.pdf.

[56] Qinqing Zheng, Amy Zhang, and Aditya Grover. Online decision transformer. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27042–27059. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zheng22c.html.

[57] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1273–1286. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/099fe6b0b444c23836c4a5d07346082b-Paper.pdf.

A Appendix

A.1 Elicitation feasibility

Assume a scenario where there are 10,000 principals and an agent with an open-ended decision space such that each principal may provide a unique perspective related to decisions the agent could make in a given month. Consider 10,000 perspectives, and assume directly eliciting a principal's view on a perspective involves casting a single vote which takes 10 seconds each. Direct elicitation of all votes equates to 28 hours of time spent voting per principal per month (280,000 hours per month in total). This is unreasonable, but theoretically possible. However, increasingly this to just 100,000 principals equates to 280 hours per principal per month – more than a full time job –, and increasing it to 1,000,000 principals equates to 2,800 hours per principal per month which is more than the number of hours in a month. There are more than 300,000,000 principals (US citizens) to the US government agent.

A.2 Scaling behavior of EIO for CDS based on previous approaches

Consider a human-perspective matrix Θ corresponding to N humans and K perspectives spanning Q different contexts (ie. elicited by Q different questions or prompts). Assume the corresponding CDS generates finite number of k perspectives for each context so $K = Qk$. The number of elements in Θ is $NQk \sim O(NQ)$. Bilich et. al. demonstrate an EIO approach to CDS in [19] where only a finite number of votes per person, c, is needed to recover an approximation of Θ for a single context. This means the number of votes which need to be sampled for a Θ spanning Q contexts is $NQc \sim O(NQ)$.

A.3 Nuclear Norm model details

The nuclear norm model likelihood is given by:

$$p(A, D, C \mid u) = \prod_{i,j \in A} \sigma(u_{ij}) \prod_{i,j \in D} (1 - \sigma(u_{ij})) \prod_{i,j,k \in C} \sigma(m_{ij} - m_{ik})$$

where $A = \{i, j | a_{ij}\}$, $D = \{i, j | d_{ij}\}$, and $C = \{i, j, k | c_{ijk}\}$. The NN model posterior is given by

$$p(u \mid A, D, C) = \frac{1}{Z} p(A, D, C \mid u) \cdot \mathbf{1}_{||u||_* < \tau}$$

$$Z := \int_{\mathbb{R}^{n \times m}} p(A, D, B \mid M) \cdot \mathbf{1}_{||M||_* < \tau} \mathrm{d}M.$$

where $\mathbf{1}_{||u||_* < \tau}$ denotes a uniform prior over the nuclear norm ball for u of radius τ. Leaning happens through optimization via SGD of the MAP estimator.

A.4 Data and training details

For each question in the dataset, there are an average of 529 perspectives, 14455 agreement exercises, and 10151 pair-choice exercises. The full dataset consists of 172,242 votes across 3,703 perspectives made by 534 humans.

In training the STUMP model, we use $L2$ regularization on the transformation weight $\mathbf{M}$ and apply dropout on the universal embedding. We use a pre-trained multilingual universal sentence encoder (Apache-2.0 license) to generate perspective embeddings $F(p_i|\phi)$. Hyper-parameters are tuned by grid search, and the following values are used in our experiments: regularization rate = 0.01, dropout rate = 0.3, latent participant embedding dimensions = 50, and learning rate = 0.001 with the Adam optimizer. The models are implemented with PyTorch, and trained on a PC with 4-core Intel i7-1165G7 processor.

A.5 Elicitation primitive experiment details

The exercises in the training sets consist of a mix of agreement and pair-choice exercises where the agreement exercises account for 60%–75% of the total exercises.

A.5.1 Type I: Single context, overlapping sample

To simulate the asynchronous CD scenario, participant's votes pertaining to perspectives submitted after their own become the validation set E_v, and votes pertaining to perspectives submitted before their own become the training pool. We modulate vote density by adjusting the fraction of votes from the training pool which are randomly partitioned into the training set E_t. To simulate the synchronous CD scenario, we randomly partition votes such that the training set E_t and validation set E_v are the same size as those in the asynchronous scenarios for each experiment.

A.5.2 Type II: Single context, non-overlapping sample

Type I to Type II transition. For each question, we randomly split the participants into two equal sized groups H_{s1} and H_{s2}, and their corresponding perspectives P_{s1} and P_{s2}. We select votes on perspectives in P_{s2} by participants in H_{s1} as the validation set, $E_v = \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s1}, j, k \in P_{s2}\}$. The train set consists of the following three pieces: votes on perspectives in P_{s1} by participants in H_{s1}, votes on perspectives in P_{s2} by participants in H_{s2}, and votes perspectives in P_{s1} submitted by a fraction p of participants in H_{s2}, where $p \in [0, 1]$, $E_t = \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s1}, j, k \in P_{s1}\} \cup \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s2}, j, k \in P_{s2}\} \cup \{e^a_{ij}, e^c_{ijk} \mid i \in H\prime_{s2}, j, k \in P_{s1}, H\prime_{s2} \subset H_{s2}, |H\prime_{s2}|/|H_{s2}| = p\}$. When $p = 1$, the train data has maximum participant overlap, i.e., it contains max number of participants with votes on perspectives in both P_{s1} and P_{s2}. When $p = 0$, the train data has zero participant overlap, i.e., it contains with votes on perspectives in both P_{s1} and P_{s2}.

Type II with Different languages. For each question, we randomly split the participants into two groups $H_{s1}$and H_{s2}, and their corresponding perspectives P_{s1} and P_{s2}. The train set consists of exercises pertaining to perspectives in P_{s1} submitted by participants in H_{s1}, and exercises pertaining to perspectives in P_{s2} submitted by participants in H_{s2}, $E_t = \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s1}, j, k \in P_{s1}\} \cup \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s2}, j, k \in P_{s2}\}$. The validation set consists of exercises pertaining to perspectives in P_{s1} submitted by participants in H_{s2}, and exercises pertaining to perspectives in P_{s2} submitted by participants in H_{s1}, $E_v = \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s2}, j, k \in P_{s1}\} \cup \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s1}, j, k \in P_{s2}\}$. The second group perspectives are translated to nine languages respectively: German (de), Spanish (es), French (fr), Italian (it), Japanese (ja), Dutch (nl), Portuguese (pt), simplified Chinese (zh) and traditional Chinese (zhtw).

A.5.3 Type III: Multiple context, overlapping sample

Transfer learning. Given a pair of contexts Q_1 and Q_2, we denote all perspectives in Q_1 as P^1 and all perspectives in Q_2 as P^2. We split the participants who participated in both contexts randomly into two groups H_{s1} and H_{s2}. The pre-train set contains exercises pertaining to perspectives in P^1 submitted by all participants, $E_{t1} = \{e^a_{ij}, e^c_{ijk} \mid i \in H, j, k \in P^1\}$. The second train set contains exercises pertaining to perspectives in P^2 submitted by participants in H_{s2}, $E_{t2} = \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s2}, j, k \in P^2\}$. The validation set contains exercises pertaining to perspectives in P^2 submitted by participants in H_{s1}, $E_v = \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s1}, j, k \in P^2\}$. Participant fraction r is defined as $r = n(H_{s2})/[n(H_{s1})+n(H_{s2})]$.

The model is pre-trained using the pre-train set E_{t1}, and the learnt participant embeddings $\mathbf{z}$ are transferred to the next training cycle using the second train set E_{t2}. In the second training cycle, $\mathbf{z}$ is fixed while the new context transformation $\mathbf{M}^2$ is learned. Then the participant embeddings $\mathbf{z}$ learnt from pre-training and context transformation $\mathbf{M}^2$ learnt from the second training are together used to make predictions for exercises in the validation set.

Increased pre-training data. Given a target context Q_t and its perspectives P^t, we select a set of $N - 1$ contexts that do not contain Q_t as the pre-train contexts. We split the participants who participated in all of the selected questions randomly into two equal sized groups H_{s1} and H_{s2}. The pre-train set contains exercises pertaining to perspectives in the pre-train contexts submitted by all participants, $E_{t1} = \cup_{n=1...N, n\neq t}\{e^a_{ij}, e^c_{ijk} \mid i \in H, j, k \in P^n\}$. The second train set contains exercises pertaining to perspectives in P^t submitted by participants in H_{s2}, $E_{t2} = \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s2}, j, k \in P^t\}$. The validation set contains exercises pertaining to perspectives in P^t submitted by participants in H_{s1}, $E_v = \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s1}, j, k \in P^t\}$.

A.5.4 Type IV: Multiple context, non-overlapping sample

Type III to Type IV transition. Given a pair of contexts Q_1 and Q_2 and their corresponding perspectives P^1 and P^2, we split the participants who participated in both contexts randomly into two equal sized groups H_{s1} and H_{s2}. The pre-train set contains exercises pertaining to perspectives in P^1 submitted by participants in H_{s1} and exercises pertaining to perspectives in P^1 submitted by a fraction p of participants in H_{s2}, where $p \in [0, 1]$, $E_{t1} = \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s1}, j, k \in P^1\} \cup \{e^a_{ij}, e^c_{ijk} \mid i \in H\prime_{s2}, j, k \in P^1, H\prime_{s2} \subset H_{s2}, |H\prime_{s2}|/|H_{s2}| = p\}$. The second train set contains exercises pertaining to perspectives in P^2 submitted by participants in H_{s2}, $E_{t2} = \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s2}, j, k \in P^2\}$. The validation set contains exercises pertaining to perspectives in P^2 submitted by participants in H_{s1}, $E_v = \{e^a_{ij}, e^c_{ijk} \mid i \in H_{s1}, j, k \in P^2\}$. When $p = 1$, E_{t1} and E_{t2} have maximum participant overlap. When p = 0, they have zero participant overlap.

A.6 Data efficiency scaling for elicitation protocols

A.6.1 Case I

The first sample elicits mc votes and initializes Θ with $N = m$ participants and $K = m$ perspectives. Each additional sample elicits an additional $2mc$ votes and adds m new participants and m new perspectives to Θ. After q samples, $n_t = mc(2q - 1)$ votes have been elicited, and Θ is comprised of $K = mq$ perspectives and $N = mq$ humans resulting in $n(\Theta) = m^2q^2$. Noting the relation $q = (n_t + mc)/(2mc)$, then $n(\Theta) = n_t^2/(4c^2) + (n_t m)/(2c) + m/4 \sim O(n_t^2)$, data leverage is $\beta = n_t/(4c^2) + m/(2c) + m/(4n_t) \sim O(n_t)$, and $\frac{\partial \beta}{\partial n_t} \approx 1/(4c^2)$ for $n_t >> \sqrt{m}/2$.

A.6.2 Case II

The first sample elicits mc votes and initializes Θ with $N = m$ participants and $K = m$ perspectives. Each additional sample elicits an additional mc votes and adds $m/2$ new participants and m new perspectives to Θ. After q samples, $n_t = mcq$ votes have been elicited, and Θ is comprised of $K = mq$ perspectives and $N = (m/2)(q + 1)$ humans resulting in $n(\Theta) = (m^2q^2 + m^2q)/2$. Noting the relation $q = n_t/(mc)$, then $n(\Theta) = n_t^2/(2c^2) + (n_t m)/(2c) \sim O(n_t^2)$, data leverage is $\beta = n_t/(2c^2) + m/(2c) \sim O(n_t)$, and $\frac{\partial \beta}{\partial n_t} = 1/(2c^2)$

A.7 Protocol simulation experiments

Case I Given a set of perspectives corresponding to Q different contexts $P^1...P^Q$, and Q equal sized non-overlapping sets of participants $H_{s1}...H_{sQ}$, we construct the following experiment for each $q = 2...Q$: let Θ^q be the HPM corresponding to $\cup_{n=1}^q\{H_{sn}\}$ and $\cup_{n=1}^q\{P^n\}$. Partition the training set in each case to correspond to local samples denoted by the shaded areas in Figure 9 such that $E_t^q = \cup_{n=2}^q\{e^a_{ij}, e^c_{ijk} \mid i \in H_{sn}, j, k \in P^n\} \cup \{e^a_{ij}, e^c_{ijk} \mid i \in \cup_{n=1}^q H_{sn}, j, k \in P^1\}$. Let the validation set E_v contain all other exercises in Θ^q, denoted by the white areas in 9.

Case II Given a set of perspectives corresponding to Q contexts $P^1...P^Q$, and Q equal sized partially overlapping sets of participants $H_{s1}...H_{sQ}$ with adjacent overlap p defined by $|H_{si} \cap H_{si+1}| = p * |H_{si}|$, we construct the following experiment for each $q = 2...Q$:let Θ^q be the HPM corresponding to $\cup_{n=1}^q\{H_{sn}\}$ and $\cup_{n=1}^q\{P^n\}$, the train set is denoted by the shaded areas in Figure 10, $E_t^q = \cup_{n=1}^q\{e^a_{ij}, e^c_{ijk} \mid i \in H_{sn}, j, k \in P^n\}$, and the validation set E_v contains all other exercises in Θ^q, denoted by the white areas 10.

A.8 Speculation on AGI alignment

We show how pre-trained LLMs support elicitation inference. It is likely inference accuracy, and preference signal quality, will improve as the capability of the LLM used increases. SOTA foundation models (FMs) are beginning to demonstrate general capabilities spanning those of LLMs and RL agents [41, 43, 55, 56, 57]. This suggests a possible scenario where an FM may be used in EIO to generate signals which steer the reward function of an agent whose decisions are made by that same FM. While such a scenario presents *significant* risks, it could also provide an opportunity to create a situation where the quality and complexity of the signals driving an AGI's alignment with a population of principals (like Humanity) increases as its capability grows. However, whether the

scaling of EIO quality would be proportional to capability in the right way to satisfy this remains an open question.

A.9 Data set

Data used for this research comes from periodic load testing done by a private company for internal quality assurance, and was fully de-identified prior to use in the research. The researchers identified that portions of it could be interesting for this work due to the scale of the data and the predictive challenges associated with the nuance and ambiguity of certain questions that were asked. Data made available for this research came from the following open-ended questions asked during the collective dialogue load test:

- *During the pandemic, what fun experiences have you newly discovered? These can be any activities, hobbies, or events you consider fun.*
- *Please tell me more: what is more important to you now?*
- *And conversely, what is less important to you now?*
- *What new routines or behaviors have you started, if any, based on what you now think is important?*
- *How or where did you learn the new routine or behavior?*
- *Tell me about a product or service you have used MORE than ever in the past year. Why did you use it more?*
- *Please tell me more – how is this impacting your decisions, if at all?*

A.10 Participant payment and consent

The 534 participants who joined the load testing exercise had been solicited through online panel providers for an incentive of approximately US$15/hour US, and consented to the policy shown in 13. While this consent form does mention PII, this was a general consent form for the data provider and no PII was known to be contained in the data set made available for this research. The researchers were structurally limited from accessing any directly collected PII due to system design, and none of raw outputs that were inspected directly during the research included any PII.

Figure 13: Screenshot of consent policy presented to all participants as part of the load test.

Conversation Terms

By clicking "**accept**" you agree to the Conversation Terms (including as to confidentiality) and to participate in the conversation subject to such terms.

PI Policy

By clicking "**accept**," you agree that is collecting/processing personal information/personal data ("**PI**") (such as demographic, employment related, consumer behaviours, etc.) on behalf of entities sponsoring this research activity ("**Sponsors**" – per *Conversation Terms*) which is used by the Sponsors to gather research insights. We may also share details of your participation with any sample provider who recruited you. For more information about how your PI is processed, your rights and s contact details, see our Privacy Policy; and with respect to a third party sample provider (if you were recruited by a third party), such third party sample provider's privacy policy applies. Your PI is collected and stored on servers located **in the USA** (hosted by Google as a subprocessor) which perform data analytics to gather research insights, such data is made available to the Sponsors and may be **transferred by the Sponsors worldwide**.

Show Less

Sensitive Data Processing

By clicking "**accept**," you agree where the Conversation seeks 'special category' or 'sensitive' PI, such as your race, religion, ethnicity, political opinions, health or sexual orientation, you agree to allow for all such PI to be collected and used to gather research insights for our Sponsors and details of your participation may be provided to your sample provider. Your PI is collected and stored on servers located **in the USA** (hosted by Google as a subprocessor) and may be **transferred by the Sponsor worldwide**.
You may withdraw your consent at any time, but this will not affect any use of your information before consent has been withdrawn. All PI will be held until the Sponsor requests deletion or until you withdraw consent.

Show Less

Cookies

We use cookies to improve your user experience and analyze website traffic, and may share your site usage data with our analytics partners and service providers, as outlined in our Privacy Policy. By clicking "**accept**," you agree cookies will be stored on your device. You can change your cookie settings at any time by clicking on the menu button in the top left corner of the Conversation.

Strictly Necessary Cookies (Always Active)

These cookies are necessary for the website to function and cannot be switched off in our systems. You can set your browser to block or alert you about these cookies, but then some parts of the site may not work. These cookies do not store any personally identifiable information.

Statistical Cookies

These cookies allow us to count visits and traffic sources so we can measure and improve the performance of our site. They help us to know which pages are the most and least popular and see how visitors move around the site. If you do not allow these cookies we will not know when you have visited our site, and will not be able to monitor its performance.

Show Less

Selections marked with * are required

Select All

- [ ] I agree to the Conversation Terms *
- [ ] I agree to the PI Policy *
- [ ] I agree to the Sensitive Data Processing *
- [ ] I consent to the use of all cookies (optional)

Accept