Neural Thurstone Model: Leveraging Latent Spaces for Collective Intelligence in Ranking Predictions

Necdet Gurkan (necdetgurkan@umsl.edu)
College of Business, University of Missouri-St. Louis
St. Louis, MO 63121, USA

Jin Bai (jbck5@umsl.edu)
College of Business, University of Missouri-St. Louis
St. Louis, MO 63121, USA

Abstract

Thurstone models have been widely applied in wisdom-of-the-crowd applications to aggregate individual rankings due to their ability to represent individual knowledge and achieve high accuracy. However, they lack the ability to generalize even across highly similar items and cannot leverage external knowledge bases or learned machine representations. In this work, we extend Thurstone models for partial ranking data by introducing a latent construct that maps pretrained vector representations to latent truths. These representations are fine-tuned through a single neural network layer, enhancing the model's ability to capture meaningful ranking structures. We evaluate our neural Thurstone model across objective ranking tasks, including animal speeds, material hardness, and the longitudinal positioning of U.S. states from west to east. Our results demonstrate that the extended model improves aggregation accuracy in sparse data settings and generalizes to novel items with moderate predictive accuracy, highlighting its potential to enhance collective intelligence in ranking-based inference.

Keywords: Wisdom of the crowd; Bayesian neural networks; Thurstone model; Ranking prediction

Introduction

Collective intelligence refers to decision-making processes that harness the collective opinions of multiple individuals, often resulting in higher-quality decisions than those made by a single person. This phenomenon, commonly known as the *wisdom of the crowd,* relies on various modes of expressing knowledge, preferences, and judgments (Surowiecki, 2005; Wallsten and Budescu, 1983). These modes include ratings, rankings, pairwise comparisons, and open-ended descriptions. Among these, ranking is particularly effective for organizing information and articulating preferences.

Rankings have been widely utilized across various fields, serving purposes in both social science and machine learning. In social sciences, ranking data is often employed to aggregate and analyze people's preferences, offering insights into collective decision-making, societal values, and group dynamics (Cohen et al., 1997; Cook and Seiford, 1978; Sühr et al., 2021). In machine learning, ranking plays an important role in applications such as search engines, where algorithms prioritize and rank search results based on relevance (Su et al., 2014), and in the development of recommendation systems, such as those used by streaming platforms to rank and suggest content based on user preferences (Karatzoglou et al., 2013).

Thurstone models have been extensively utilized to harness collective intelligence across various domains, including aggregating factual knowledge (Lee et al., 2014), generating predictions (Selker et al., 2017), and modeling preferences (Lee and Ke, 2022). These models offer a probabilistic framework for analyzing the cognitive processes underlying human judgments, allowing for the extraction of latent orders from ranking data and facilitating the inference of preferences, expertise, and knowledge (Lee et al., 2012; Steyvers et al., 2009; Thurstone, 1994).

Many applications of Thurstone models in these domains rely on observed item rankings. While extensions of these models have been developed to accommodate partial rankings (Montgomery et al., 2024), they still depend on at least one observed ranking for each item. Consequently, the lack of generalizability to unseen items presents a significant limitation, particularly in scenarios where no direct observations exist for certain items. Examples include predicting preferences for new products or understanding opinions about previously unconsidered concepts. This limitation constrains the performance of Thurstone models in sparse data settings and creates a cold-start problem, where the models struggle to generalize to new, unseen items—even those closely related to previously ranked ones.

Moreover, the Thurstone models are unable to leverage existing structured knowledge or learned machine representations that could provide valuable information about known entities and their relationships. These limitations restrict its applicability in complex or dynamic contexts where the interrelatedness of judgments and the availability of preexisting knowledge play pivotal roles.

Modern machine-learning methods address many of these issues by enabling the prediction of out-of-sample items through models that learn over latent representations. Recent advancements in deep learning can process vast quantities of data from diverse modalities, including text, images, and audio, to identify patterns and relationships that generalize beyond the training data (LeCun et al., 2015). These techniques have enabled the creation of expressive high-dimensional vector representations for words, sentences, visual scenes, and objects.

As these vector representations are argued to approximate human mental representations (Piantadosi et al., 2024), they have been effectively used as inputs to linear models for

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

predicting individual and aggregate evaluations across various domains, such as perceived risk (Bhatia, 2019), impression formation (Peterson et al., 2022), leadership perceptions (Bhatia et al., 2022), and creative writing assessments (Johnson et al., 2022). Additionally, these representations have been integrated into psychometric models as a latent construct to map them to consensus beliefs (Gürkan and Suchow, 2023), demonstrating the potential for bridging pretrained representations with collective intelligence models.

In this paper, we extend Thurstone models for partial ranking data by introducing a latent construct that maps pretrained vector representations to latent truths. These representations are fine-tuned through a single neural network layer, improving the model's ability to capture meaningful ranking structures. We evaluate our neural Thurstone model across objective ranking tasks, including animal speeds, material hardness, and the longitudinal positioning of U.S. states from west to east, where rankings are based on measurable criteria.

Additionally, we examine the potential of LLMs as proxies for human cognition in ranking tasks by incorporating a latent construct derived from LLM-generated reasoning. While LLMs like ChatGPT have limitations, they offer a novel approach to understanding the underlying mechanisms of human ranking judgments (Bhatia, 2024). By leveraging their reasoning capabilities, we explore whether these models can enhance ranking predictions and provide meaningful insights into decision-making processes.

The structure of this paper is as follows. First, we review pretrained vector representations and their applications. Next, we introduce the Thurstone model and the mathematical formalism that underpins our extended approach. We then describe the datasets and methodologies used for validation. Finally, we present the results from applying our model across multiple domains and discuss their implications.

Machine Representations

Recent advancements in deep learning have made it possible to generate expressive high-dimensional vector representations, commonly referred to as embeddings, for objects, entities, and concepts (LeCun et al., 2015). Deep learning models process input data through multiple hidden layers, where each layer transforms the data into intermediate representations. These learned embeddings capture meaningful relationships, patterns, and similarities within the data, ultimately influencing the model's predictions in the output layer (e.g., classification labels, actions, or ratings). By fine-tuning these embeddings during training, deep learning models create structured representations that effectively map input data to meaningful responses (LeCun et al., 2015).

In natural language processing (NLP), distributional semantics provides a framework for deriving embeddings from large-scale textual data (Boleda, 2020). These embeddings are based on the principle that word meaning emerges from co-occurrence patterns in language (Landauer and Dumais, 1997). Words frequently used in similar contexts tend to have closer representations in high-dimensional semantic spaces. Pre-trained embeddings, such as word2vec (Church, 2017) and BERT (Acheampong et al., 2021), serve as inputs for more sophisticated models that are fine-tuned with participant data to approximate human knowledge structures (Bhatia et al., 2019).

Word embeddings have been applied to a variety of tasks, including predicting risk perception (Bhatia, 2019), evaluating leadership impressions (Bhatia et al., 2022), and modeling biases in social judgments (Caliskan et al., 2017). Additionally, deep learning advancements have enabled the generation of sentence embeddings that capture broader contextual meaning, supporting research in commonsense reasoning (Forbes et al., 2019), relational knowledge (Bouraoui et al., 2020), and cognitive diversity estimation (Gurkan and Yan, 2023).

Similarly, deep neural networks trained on large-scale image datasets can classify, detect, and recognize objects and concepts, learning intermediate representations that generalize effectively across tasks. Researchers have leveraged these embeddings to study human perceptual judgments, such as object memorability (Khosla et al., 2015) and typicality (Lake et al., 2015). While machine-derived representations are highly informative, they do not fully capture human psychological structures. To bridge this gap, researchers have integrated these embeddings with psychological models to examine the alignment between computational and human cognitive representations (Peterson et al., 2022; Sucholutsky and Griffiths, 2024).

Thus, pretrained intermediate representations provide structured, high-dimensional features that capture underlying semantic, perceptual, or conceptual relationships, making them valuable as knowledge representations in computational models. By incorporating these embeddings into computational frameworks, researchers can model human judgments, preferences, and expertise with greater precision. For example, embeddings derived from LLMs and image-based models can be used to estimate consensus in subjective evaluations (Gurkan and Suchow, 2022; Gürkan and Suchow, 2023). In the context of Thurstone models, these embeddings offer a means to extend traditional ranking-based inference by integrating rich, data-driven representations, enhancing the ability to infer latent structures in human judgments.

Thurstone Model

The Thurstone model was originally developed to study the law of comparative judgments and consensus opinions derived from rank data by representing item characteristics as latent psychological dimensions along a continuum (Luce, 1994; Thurstone, 1927). This framework has since been extended to wisdom-of-the-crowd applications, leveraging its ability to model collective judgments (Lee, 2024). The central premise of these applications is that, for many tasks, the group's central tendency often yields a highly accurate outcome (Galton, 1907). In this context, the collective re-

sponse serves as a proxy for the actual answer, with individuals whose judgments align more closely with the group's central tendency considered to possess greater knowledge or expertise. Similarly, the Thurstone model maps the latent dimension to the criterion of interest, item positions along this dimension to the ground truth, and the variability in these representations to the uncertainty individuals experience regarding the true value (Lee et al., 2014).

Specifically, the model assumes that individuals generate rankings by sampling from mental distributions based on the latent positions of items, aligning naturally with the psychological principles underlying Thurstone models (Marden, 1996). The Thurstone model is defined by two key parameters: μ and σ. The parameter μ represents the underlying latent truth of each item, indicating its position on the psychological continuum, while σ reflects the precision of each individual, capturing their level of knowledge or accuracy with respect to the criterion (Lee et al., 2014).

In the Thurstone model, individuals are assumed to draw mental samples x_{i_1}, x_{i_2}, . . ., from a Gaussian distribution centered on the ground truths μ_{j_1}, μ_{j_2} . . ., with a standard deviation determined by σ_i, which represents the individual's precision. Each mental sample is thus drawn as:

$$x_{ij} = \text{Gaussian}(\mu_j, \frac{1}{\sigma_i^2}), \tag{1}$$

where the Gaussian distribution is parameterized by its mean (μ_j) and precision ($\frac{1}{\sigma_i^2}$). As the precision increases, the mental samples become more tightly centered around the ground truth, reflecting a higher level of knowledge or expertise. The precision parameter in the Thurstone model has been shown to correlate closely with expertise levels, making it a critical component in understanding individual differences in judgment accuracy (Lee, 2024; Lee et al., 2014).

Neural Thurstone Model

While the Thurstone model provides a robust computational framework for harnessing collective intelligence in rank-order judgments, it is not well-suited for integrating recent advancements in machine learning. The Thurstone model represents the structure of individual and collective judgments as a probabilistic distribution over a latent psychological continuum, where item positions correspond to their inferred true ranking. In this representation, ranked items lack intrinsic structure and are only connected through correlations across participants.

However, this computational framework has several limitations. First, it treats each ranked item independently, preventing information about one item from informing our understanding of others. Second, it requires a sufficient number of ranked items to accurately characterize the latent rank order, making it less effective in sparse data settings. Finally, it does not leverage structured knowledge bases that encode information about known entities and their relationships, limiting its ability to generalize beyond observed rankings.

To overcome these challenges, we extend the Thurstone model to provide a more data-driven representation of a latent psychological continuum than traditional probabilistic ranking models that rely solely on observed rankings. In this extension, we define a latent construct (Gurkan and Suchow, 2022; van den Bergh et al., 2020), a process that maps an item to the latent psychological continuum through an intermediate representation. In this work, we use pretrained deep neural network representations, or embeddings, as latent constructs. We then fine-tune these representations by passing them through a single neural network layer.

This layer consists of two learnable parameter sets: (1) θ, which transforms the pretrained embedding ϕ_j into a task-adapted representation ψ_j, and (2) ω, which maps this adapted representation to a scalar latent value μ_j indicating the item's position on the latent continuum. Both θ and ω are updated during model training via gradient-based inference. This setup allows the model to flexibly adapt pretrained knowledge to the structure of observed ranking data.

We opted for a neural network structure rather than a linear model due to its ability to capture non-linear relationships between item representations and latent rankings. While linear models assume a fixed relationship between embeddings and rank positions, neural networks can learn flexible transformations, allowing for better generalization across items with complex dependencies. This structure is particularly advantageous when dealing with sparse ranking data, where linear models may struggle to extrapolate meaningful relationships. Additionally, the non-linear activation function introduces a level of expressiveness that better aligns with human cognitive processes, which are unlikely to follow strictly linear patterns in ranking judgments.

Thus, the underlying latent truth of each item, μ_j, is obtained from:

$$\begin{aligned} \mu_j &= \psi_j \omega^T && \text{Latent truth} \\ \psi_j &= \tanh(\phi_j \cdot \theta) && \text{Fine-tuned embedding representation} \\ \theta &\sim \text{Normal}(0,1) && \text{Layer weights} \end{aligned}$$

where ψ_j is the fine-tuned embedding representation of an item. ϕ_j is its pretrained embedding, and ω is a linear transformation that maps the fine-tuned representation to the latent truth. The weight parameter θ is drawn from a standard normal distribution to introduce flexibility in the transformation process.

Methods

Data

We validated our neural Thurstone model on objective ranking tasks, where rankings are based on measurable criteria. We used datasets ranking U.S. states from west to east (e.g., Oregon, Maine), hardness of materials (e.g., diamond, steel), and animal speeds (e.g., cheetah, dove). These datasets

were obtained from Lee et al. (2014) and Montgomery et al. (2024).

The datasets varied in ranking completeness. The animal speeds datasets contained partial ranking data, where not every item was ranked by every participant. In contrast, the U.S. states from west to east, and hardness of materials datasets contained full ranking data, where all participants ranked every item.

To obtain pretrained embeddings, we used Sentence-BERT (SBERT) (Reimers, 2019), a pretrained sentence transformer, to generate a 384-dimensional vector representation for each item. These embeddings served as inputs to the latent construct in our model, allowing it to learn structured relationships between ranked items.

LLM-Generated Reasoning: Recent advancements in LLMs, such as ChatGPT, offer new opportunities to assess whether their reasoning can serve as a useful proxy for human cognition and improve decision-making processes. For instance, LLM-generated explanations of various risk sources can be compared with human rationales to evaluate their alignment (Bhatia, 2024). By analyzing the extent to which LLMs capture the underlying cognitive mechanisms that drive ranking behavior, we can determine their potential for improving ranking-based inference and decision-making models.

To examine whether incorporating LLM-generated reasoning as a latent construct enhances model performance, we generated explanations using GPT-4 for datasets with the following prompt: *"What are three reasons someone might consider {animal, material, state} to be {fast/slow, hard/soft, located in the west/east}? List three reasons for each."* We used the OpenAI API instead of ChatGPT's interactive interface, as the API allowed for automated, structured querying and response recording across multiple materials. To prevent the model from simply learning the correct rankings, we ensured that the generated reasons did not explicitly state the true rank order of the items. Instead, the responses provided general, plausible justifications based on commonly known attributes. Each generated reason was then embedded using SBERT, and the resulting embeddings were averaged to form a latent construct. In our neural Thurstone model, these averaged embeddings replaced the original item name embeddings, enabling the model to incorporate reasoning-based representations rather than relying solely on item labels.

Methods Summary

The neural Thurstone model was implemented in NumPyro (Phan et al., 2019) using the JAX backend (Bradbury et al., 2018). Model components were integrated into a single likelihood function with a set of prior distributions. Inference was performed using a Gibbs sampler (Liu, 1996) combined with the No-U-Turn Sampler (NUTS) (Hoffman, Gelman, et al., 2014), a standard Markov Chain Monte Carlo (MCMC) algorithm available in NumPyro. We ran a single chain with 10,000 warm-up iterations followed by 10,000 posterior samples. Convergence was verified by ensuring no divergent transitions and that the Gelman-Rubin diagnostic ($R_{hat} < 1.01$) indicated proper mixing.

Results

We evaluated our model using multiple validation and prediction approaches. First, we assessed whether our model could recover rankings similar to the traditional Thurstone model or improve aggregation accuracy. Second, we tested its predictive ability using a leave-one-out (LOO) validation approach, allowing direct comparison with the traditional Thurstone model. In both validation and prediction tasks, we examined whether incorporating LLM-generated reasoning enhanced aggregation and prediction accuracy, as our neural Thurstone model enables the Thurstone framework to leverage external knowledge structures. Finally, we evaluated the model's generalization ability by predicting the rankings of a novel set of items, where we randomly selected items that were not present in the dataset.

Model Validation We evaluated whether the neural Thurstone model improves the aggregation of participant rankings in objective domains or if it recovers similar rank orders as the base model. To assess this, we compared the aggregated predictions of the traditional Thurstone model with those of the neural Thurstone model. By evaluating their performance across objective ranking tasks, we aimed to determine whether incorporating pretrained embeddings enhances the model's ability to infer latent rank orders from participant responses.

Model performance was assessed using Kendall's τ distance, a rank correlation measure that quantifies agreement between predicted and actual rankings by counting the number of adjacent pairwise swaps required to align the two orders (Lee et al., 2014). A lower Kendall's τ distance indicates greater alignment between model-inferred and ground truth rankings, reflecting improved aggregation accuracy.

The aggregate rankings inferred by our neural Thurstone model outperformed the traditional Thurstone model in predicting animal speeds. Based on the ordering of posterior means for μ, the neural model achieved a lower Kendall's τ distance to the ground truth. The mean of the posterior distribution for τ under the neural model was 6, with a 95% credible interval of (4, 8), compared to a mean of 8 with a 95% credible interval of (5, 11) under the traditional model. This suggests that the neural model more reliably captures the true item ordering. The improvement is likely due to the sparsity of the dataset, where semantic relationships embedded in the pretrained features help bridge gaps in the observed rankings.

For the hardness of materials dataset, both models produced identical results, with the posterior mean ordering yielding a Kendall's τ distance of 12 and a 95% credible interval of (7, 17). Similarly, in the U.S. states from west to east dataset, the neural model demonstrated equivalent performance to the traditional model, with both models achieving a Kendall's τ distance of 7 and a 95% credible interval

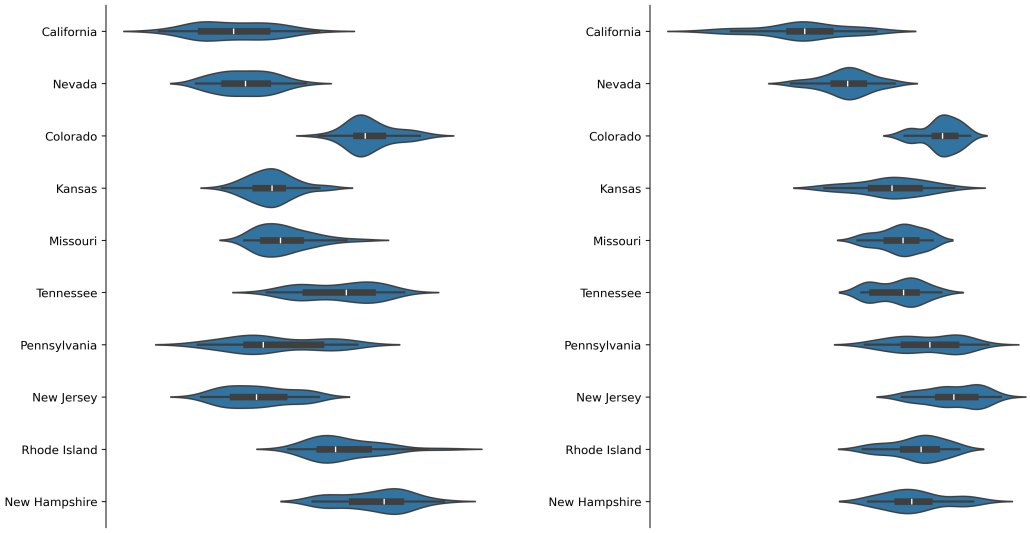

Figure 1: Neural Thurstone model predictions for novel items. States are ordered from highest to lowest based on predicted rankings, with the top representing the highest-ranked state. The predicted order is determined by the mean of the posterior distributions for the location parameter μ, visualized using violin plots. The **left** figure shows model predictions using item embeddings as latent constructs ($\tau = 14$), while the **right** figure presents predictions using LLM-generated reason embeddings as latent constructs ($\tau = 12$).

of (3, 12). These results suggest that when sufficient ranking information is available, pretrained embeddings offer no additional benefit over traditional inference.

To evaluate the impact of LLM-generated reasoning as latent constructs, we incorporated embeddings derived from GPT-4's explanations into the neural Thurstone model. The model's performance remained unchanged for the animal speed and hardness of materials datasets. However, for the U.S. states from west to east ranking, incorporating LLM-based reasoning improved the Kendall's τ distance from 7 (95% CI: 3, 12) to 6 (95% CI: 2, 10), indicating a modest gain in aggregation accuracy. This suggests that LLM-generated explanations can provide useful structured knowledge in spatial or geographically grounded ranking tasks.

Model Prediction Next, we tested whether our model could accurately predict an item's ground truth position by holding out one item and estimating its actual rank based on objective criteria. To evaluate the model's predictive accuracy, we employed a LOO validation approach. In this setup, one item was excluded during training, and its rank was predicted relative to the remaining items. This process was repeated for each item in the dataset.

Since inferred and predicted μ values are relative to the training set in each LOO iteration, direct comparisons across iterations are not meaningful. Instead, we converted each inferred μ value into rank positions within that specific LOO iteration, ensuring consistency in ranking comparisons. This approach allows us to assess how well the model predicts an unseen item's rank when trained on the remaining dataset and compare its performance to the traditional Thurstone model, which assumes all items are observed.

Rather than using Kendall's τ distance, which counts the number of adjacent pairwise swaps needed to align two rankings, we used Kendall's τ correlation, which measures the overall rank agreement while allowing for ties. This distinction is critical because, in LOO settings, the model could assign the same position to multiple items, making swap-based distance metrics ill-defined. In contrast, Kendall's τ correlation provides a more appropriate evaluation by quantifying the relative agreement between predicted and ground truth rankings, even when items have identical predicted positions.

The model's predictive performance varied across datasets. For the U.S. states from west to east ranking task, our model achieved a Kendall's τ of 0.72, outperforming the traditional Thurstone model, which had a τ of 0.68. In the hardness of materials dataset, the model performed worse than the traditional Thurstone model, with a τ of 0.12 compared to 0.38 for the traditional model. Similarly, in the animal speed ranking task, the model achieved a τ of 0.54, while the traditional Thurstone model demonstrated higher agreement with the ground truth, attaining a τ of 0.73.

To further assess whether incorporating LLM-generated reasoning as latent constructs enhances predictive performance, we applied this approach to the west-to-east ranking dataset in the LOO setting. The results showed that integrating LLM-derived embeddings improved Kendall's τ from 0.72 to 0.75, reinforcing the observation that LLM-generated knowledge may be particularly beneficial in spatial and geographically structured ranking tasks. However, for the hardness of materials and animal speed datasets, LLM-based reasoning did not improve model predictions.

Out-of-Distribution Performance To further evaluate the generalization capability of our neural Thurstone model, we conducted an additional test by randomly selecting 10 U.S. states that were not included in the original west-to-east ranking dataset. This dataset was chosen because our model's predictions had outperformed the traditional model in previous evaluations, suggesting that it effectively captures spatial relationships.

For each unseen state, we predicted its rank order based on the model's learned representations and compared these predictions to the ground truth rankings. The model using item embeddings achieved a Kendall's τ distance of 14, with a 95% credible interval of (10, 18), indicating a moderate level of agreement with the true rankings (Figure 1, left plot).

We also conducted a similar prediction task using LLM-generated reasoning as latent constructs. The model trained on reasoning-based embeddings predicted the rank order of unseen states, achieving a Kendall's τ distance of 12, with a 95% credible interval of (8, 16) (Figure 1, right plot). This result suggests that incorporating LLM-generated reasoning slightly improved the model's ability to generalize to novel items. The performance gain may be due to LLM-derived embeddings encoding richer contextual and relational information, enabling the model to make more informed ranking predictions than with item embeddings alone.

Discussion

The Thurstone model is widely used to harness collective intelligence in ranking tasks. However, it lacks the ability to make predictions on out-of-sample data and does not leverage existing structured knowledge or pretrained representations. To address these limitations and enhance its applicability in modern settings, we introduced the neural Thurstone model, which extends the traditional Thurstone framework by incorporating a latent construct that maps pretrained embeddings to latent truths.

Our results demonstrate that the neural Thurstone model slightly improves aggregation performance on animal speed rankings while outperforming human participants in the LOO prediction task for the longitudinal positioning of U.S. states. Additionally, we showed that the model can generalize to novel items, demonstrating moderate predictive accuracy in out-of-distribution ranking tasks. These findings suggest that incorporating pretrained embeddings and external knowledge representations into Thurstone models enhances their ability to infer latent rank structures and predict unseen rankings, making them more adaptable to modern machine learning applications.

Future extensions

While pretrained embeddings effectively capture semantic relationships among items, they lack interpretability. To address this limitation, the Thurstone model can be extended by incorporating interpretable features as latent constructs, mapping them to latent truths through a regression framework. This approach would enable researchers to examine feature importance in ranking decisions by analyzing the corresponding regression coefficients. For instance, in subjective ranking tasks, researchers could identify the most influential attributes shaping perceived leadership or presidential rankings.

The Thurstone model has been extended to accommodate multiple latent rankings for subjective evaluations, where no objective ground truth exists (Selker et al., 2017). However, this framework can be further enhanced using Bayesian nonparametric methods, which offer greater flexibility in modeling both individual- and group-level variations. Such an approach would enable an end-to-end probabilistic framework, allowing the model to dynamically infer the number of latent ranking structures and adapt to complex, heterogeneous ranking behaviors across different domains.

Limitations

One key limitation of our approach is the use of LLM-generated reasons as a latent construct without direct validation against human-generated explanations. While LLM-derived embeddings capture semantic relationships and can enhance ranking predictions, their utility is particularly promising for subjective evaluations, where understanding the rationale behind preferences is crucial. By incorporating reasoning-based embeddings, we can move beyond purely numerical rankings and gain insights into the factors driving human judgments.

However, to rigorously assess whether LLM-generated reasons align with human reasoning, it is necessary to collect human-provided justifications for rankings. Future work should involve vector space analysis to compare semantic similarity between LLM and human explanations, as well as coefficient analysis within a regression framework to determine the similarities between LLM- and human-generated reasons. This would enable a deeper understanding of how well LLM-generated rationales reflect human cognition and whether they can serve as a reliable proxy for human decision-making processes in ranking tasks.

Additionally, our model's generalization performance is constrained by the number of items being ranked in the datasets. Since ranking models rely on relative comparisons, a dataset with a limited number of items may not provide sufficient variability for the model to accurately generalize across unseen rankings. Expanding the datasets to include a larger set of ranked items would allow the model to capture broader ranking structures, improving its ability to infer latent rank orders and make more accurate predictions on out-of-sample items. Future work should explore how dataset size and diversity affect the model's performance, particularly in subjective ranking tasks, where ranking distributions may be more variable across individuals.

References

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of bert-based approaches. *Artificial Intelligence Review*, *54*(8), 5789–5829.

Bhatia, S. (2019). Predicting risk perception: New insights from data science. *Management Science*, *65*(8), 3800–3823.

Bhatia, S. (2024). Exploring variability in risk taking with large language models. *Journal of Experimental Psychology: General*.

Bhatia, S., Olivola, C. Y., Bhatia, N., & Ameen, A. (2022). Predicting leadership perception with large-scale natural language data. *The Leadership Quarterly*, *33*(5), 101535.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, *6*(1), 213–234.

Bouraoui, Z., Camacho-Collados, J., & Schockaert, S. (2020). Inducing relational knowledge from bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 7456–7463.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., et al. (2018). Jax: Composable transformations of python+ numpy programs.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Church, K. W. (2017). Word2vec. *Natural Language Engineering*, *23*(1), 155–162.

Cohen, W. W., Schapire, R. E., & Singer, Y. (1997). Learning to order things. *Advances in Neural Information Processing Systems*, *10*.

Cook, W. D., & Seiford, L. M. (1978). Priority ranking and consensus formation. *Management Science*, *24*(16), 1721–1732.

Forbes, M., Holtzman, A., & Choi, Y. (2019). Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*.

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451.

Gurkan, N., & Suchow, J. (2022). Cultural alignment of machine-vision representations. *SVRHM 2022 Workshop@ NeurIPS*.

Gurkan, N., & Yan, B. (2023). Chatbot catalysts: Improving team decision-making through cognitive diversity and information elaboration. *ICIS 2023 Proceedings*, *18*.

Gürkan, N., & Suchow, J. W. (2023). Harnessing collective intelligence under a lack of cultural consensus. *arXiv preprint arXiv:2309.09787*.

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623.

Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., van Hell, J., Kennedy, E., Sullivan, G. F., Taylor, C. L., et al. (2022). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, 1–34.

Karatzoglou, A., Baltrunas, L., & Shi, Y. (2013). Learning to rank for recommender systems. *Proceedings of the 7th ACM Conference on Recommender Systems*, 493–494.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. *Proceedings of the IEEE International Conference on Computer Vision*, 2390–2398.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, M. D. (2024). Using cognitive models to improve the wisdom of the crowd. *Current Directions in Psychological Science*, *33*(5), 308–316.

Lee, M. D., & Ke, M. Y. (2022). Modeling individual differences in beliefs and opinions using thurstonian models. *The Cognitive Science of Belief: A Multidisciplinary Approach*, 488.

Lee, M. D., Steyvers, M., De Young, M., & Miller, B. (2012). Inferring expertise in knowledge and prediction ranking tasks. *Topics in Cognitive Science*, *4*(1), 151–163.

Lee, M. D., Steyvers, M., & Miller, B. (2014). A cognitive model for aggregating people's rankings. *PloS One*, *9*(5), e96431.

Liu, J. S. (1996). Peskun's theorem and a modified discrete-state gibbs sampler. *Biometrika*, *83*(3).

Luce, R. D. (1994). Thurstone and sensory scaling: Then and now. *Psychological Review*, *34*, 271–277.

Marden, J. I. (1996). *Analyzing and modeling rank data*. CRC Press.

Montgomery, L. E., Bradford, N., & Lee, M. D. (2024). The wisdom of the crowd with partial rankings: A bayesian approach implementing the thurstone model in jags. *Behavior Research Methods*, *56*(7), 8091–8104.

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, *119*(17), e2115228119.

Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*.

Piantadosi, S. T., Muller, D. C., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2024). Why

concepts are (probably) vectors. *Trends in Cognitive Sciences*, *28*(9), 844–856.

Reimers, N. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Selker, R., Lee, M. D., & Iyer, R. (2017). Thurstonian cognitive models for aggregating top-n lists. *Decision*, *4*(2), 87.

Steyvers, M., Miller, B., Hemmer, P., & Lee, M. (2009). The wisdom of crowds in the recollection of order information. *Advances in Neural Information Processing systems*, *22*.

Su, A.-J., Hu, Y. C., Kuzmanovic, A., & Koh, C.-K. (2014). How to improve your search engine ranking: Myths and reality. *ACM Transactions on the Web (TWEB)*, *8*(2), 1–25.

Sucholutsky, I., & Griffiths, T. (2024). Alignment with human representations supports robust few-shot learning. *Advances in Neural Information Processing Systems*, *36*.

Sühr, T., Hilgard, S., & Lakkaraju, H. (2021). Does fair ranking improve minority outcomes? understanding the interplay of human and algorithmic biases in online hiring. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 989–999.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Thurstone, L. L. (1994). A law of comparative judgment. *Psychological Review*, *101*(2), 266.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

van den Bergh, D., Bogaerts, S., Spreen, M., Flohr, R., Vandekerckhove, J., Batchelder, W. H., & Wagenmakers, E.-J. (2020). Cultural consensus theory for the evaluation of patients' mental health scores in forensic psychiatric hospitals. *Journal of Mathematical Psychology*, *98*, 102383.

Wallsten, T. S., & Budescu, D. V. (1983). State of the art—encoding subjective probabilities: A psychological and psychometric review. *Management Science*, *29*(2), 151–173.