Leveraging a Cognitive Model to Measure Subjective Similarity of Human and GPT-4 Written Content

Tyler Malloy and **Maria José Ferreira** and **Fei Fang** and **Cleotilde Gonzalez**
tylerjmalloy@cmu.edu mariajor@andrew.cmu.edu feifang@cmu.edu coty@cmu.edu
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh PA, USA

Abstract

Cosine similarity between two documents can be computed using token embeddings formed by Large Language Models (LLMs) such as GPT-4, and used to categorize those documents across a range of uses. However, these similarities are ultimately dependent on the corpora used to train these LLMs, and may not reflect subjective similarity of individuals or how their biases and constraints impact similarity metrics. This lack of cognitively-aware personalization of similarity metrics can be particularly problematic in educational and recommendation settings where there is a limited number of individual judgements of category or preference, and biases can be particularly relevant. To address this, we rely on an integration of an Instance-Based Learning (IBL) cognitive model with LLM embeddings to develop the Instance-Based Individualized Similarity (IBIS) metric. This similarity metric is beneficial in that it takes into account individual biases and constraints in a manner that is grounded in the cognitive mechanisms of decision making. To evaluate the IBIS metric, we also introduce a dataset of human categorizations of emails as being either dangerous (phishing) or safe (ham). This dataset is used to demonstrate the benefits of leveraging a cognitive model to measure the subjective similarity of human participants in an educational setting.

1 Introduction

When humans categorize textual information, such as when giving recommendations or learning to categorize documents, we often use our personal subjective concepts to complete the task. One example of this is giving a recommendation of a funny book to a fiend, which requires not only our own subjective conceptualization of humor, but also an understanding of the similarities and differences between ourselves and our friends. While humans perform this task with relative ease, recommendation systems (Ko et al., 2022) and educational tools (Nafea et al., 2019) typically do not have personalized measurements of subjective concepts (Gazdar and Hidri, 2020), potentially hindering their efficacy (Pal et al., 2024).

When these systems incorporate data from human judgements to determine subjective similarity, they typically do so by pooling together as many judgements from different people as they can, and aggregate their measurement (Xia et al., 2015). This approach relies on machine learning based methods (Shojaei and Saneifar, 2021), which can be effective from a machine learning perspective, since more data can mean improved document similarity metrics on average over large datasets (Kusner et al., 2015). Focusing on individuals annotations of documents has been explored in the context of domain specific knowledge such as biomedical research papers (Brown and Zhou, 2019), or for specific context like document summarizing (Zhang et al., 2003).

However to date little attention has been given to the notion of individualized metrics of similarity that account for biases and constraints specifically, which are highly relevant for educational contexts (Chew and Cerbin, 2021). Related to the domain of this work in particular, recent work has demonstrated a broad range of human opinions and levels of trust associated with cybersecurity concepts such as Trusted Execution Environments (Carreira et al., 2024). This highlights the need for individualized metrics that take into account experience in training tasks such as the anti-phishing training dataset used in this work.

In this work, we propose a method for providing personalized metrics of subjective concepts that can determine the similarity between sets of text, with additional applications in selecting educational examples and providing natural language feedback. This is done by leveraging a cognitive model of human learning and decision making that can act as a digital twin to individuals, and predict their behav-

ior and opinions on a wider set of stimuli. We focus specifically on students categorizing emails as being safe (ham) or dangerous (phishing) in a training setting to help users identify and defend against phishing email attacks. Our proposed method for providing personalized similarity metrics of documents is compared to alternative methods using a dataset of a phishing education task experiment that we additionally present in this work.

The dataset of human annotations of emails as being either ham or phishing is described in (Malloy et al., 2024) and was made publicly available on OSF[1]. This dataset consists of human annotations of email documents that are either written by cybersecurity experts or a GPT-4 model, the emails shown to participants, and conversations between human participants and a GPT-4o model providing feedback to students. In total this dataset represents 39230 human judgements from 433 participants making decisions while observing a set from 1440 GPT-4 or human generated emails, as well as 20487 messages between human participants and the GPT-4o teacher model.

This type of learning task represents a serious challenge for traditional methods of adjusting document embedding similarity metrics to conform to human behavior, such as embedding pruning (Manrique et al., 2023) or embedding weighting (Onan, 2021). This is because these approaches typically rely on a large amount of annotations collected from many participants who are expected to have the same knowledge level throughout the annotation process. Instead, we are interested in measuring the subjective similarity of documents as participants learn the document annotation task in a training setting. To do this, we employ a cognitive model that can predict the learning trajectory of each individual participant as they learn to correctly annotate these documents.

2 Background: Cognitive Model

The cognitive model used in this work to predict the subjective similarity of human participants decisions on unseen emails relies on Instance Based Learning Theory (IBLT) (Gonzalez et al., 2003). One of the benefits of employing IBL models over alternatives like Reinforcement Learning is that they base their predictions on the full history of participant experience as well as the impact that limitations like memory size and decay can have on decision making.

IBL models have been applied onto predicting human behavior in dynamic decision making tasks, including binary choice tasks (Gonzalez and Dutt, 2011; Lejarraga et al., 2012), theory of mind applications (Nguyen and Gonzalez, 2022), and practical applications such as identifying phishing emails (Cranford et al., 2019; Malloy and Gonzalez, 2024), cyber defense (Cranford et al., 2020), and cyber attack decision-making (Aggarwal et al., 2022).

2.1 Activation

IBL models work by storing instances i in memory $\mathcal{M}$, composed of utility outcomes u_i and options k composed of features j in the set of features $\mathcal{F}$ of environmental decision alternatives. These options are observed in an order represented by the time step t, and the time step that an instance occurred in is given $\mathcal{T}(i)$. Option values are determined by selecting the action that maximizes the blended value $\mathcal{V}_k(t)$. In calculating this activation, the similarity between instances in memory and the current instance is represented by summing over all attributes the value S_{ij}, which is the similarity of attribute j of instance i to the current state. This gives the activation equation as:

$$A_i(t) = \ln\left(\sum_{t' \in \mathcal{T}_i(t)} (t - t')^{-d}\right) + \mu \sum_{j \in \mathcal{F}} \omega_j (S_{ij} - 1) + \sigma\xi \quad (1)$$

The parameters that are set either by modelers or set to default values are the decay parameter d; the mismatch penalty μ; the attribute weight of each j feature ω_j; and the noise parameter σ. The default values for these parameters are $(d, \mu, \omega_j, \sigma) = (0.5, 1, 1, 0.25)$. The value ξ is drawn from a normal distribution $\mathcal{N}(-1, 1)$ and multiplied by the noise parameter σ to add random noise to the activation.

2.2 Similarity Measure

The definition of the similarity measure S_{ij} is highly influential in the behavior of the IBL model, as it determines which instances from memory are drawn from to predict utility. In simple binary choice tasks without attributes (Gonzalez and Dutt, 2011; Lejarraga et al., 2012), the similarity metric can be defined as the equality function $S_{ij} = 1$ if $i == j$ else 0. In more complex domains such

[1] https://osf.io/wbg3r/

as the phishing email identification task used in this work, one approach is to use the embeddings of emails to compare the similarity of instances, and rely on the cosine similarity metric to compute the similarity of instances in memory (Malloy and Gonzalez, 2024). The model presented in this work relies on an initial baseline similarity metric, the standard cosine similarity, to then build more individual specific metrics of similarity.

2.3 Probability of Retrieval

The probability of retrieval represents the probability that a single instance in memory will be retrieved when estimating the value associated with an option. To calculate this probability of retrieval, IBL models apply a weighted soft-max function onto the memory instance activation values $A_i(t)$ giving the equation:

$$P_i(t) = \frac{\exp A_i(t)/\tau}{\sum_{i' \in \mathcal{M}_k} \exp A_{i'}(t)/\tau} \tag{2}$$

The parameter that is either set by modelers or set to its default value is the temperature parameter τ, which controls the uniformity of the probability distribution defined by this soft-max equation. The default value for this parameter is $\tau = \sigma\sqrt{2}$.

2.4 Blended Value

The blended value determines the ultimate action selected by the model and is calculated of an option k at time step t according to the utility outcomes u_i weighted by the probability of retrieval of that instance P_i and summing over all instances in memory $\mathcal{M}_k$ to give the equation:

$$V_k(t) = \sum_{i \in \mathcal{M}_k} P_i(t) u_i \tag{3}$$

These blended values are used to determine the action a_{t+1} selected by the model at the next time step.

$$a_{t+1} = \max_{k \in K} V_k(t) \tag{4}$$

In standard IBL models, this action can be used in simulations to allow the model to gain experience in a given task. In model tracing, which is used in the method proposed in this work, the memory of instances is made up of the past observations and decisions of the participant, with the action representing a prediction of their future behavior.

3 Phishing Email Categorization Dataset

The first component of this dataset is human behavioral experiment data from a study on human categorization of emails. This experiment compared human document annotation when categorizing emails as phishing (dangerous) or ham (safe). The conditions of this experiment varied depending on the email author (Human or GPT-4) and style (plain-text or GPT-4 stylized). There was also a comparison of the method of selecting emails to show to participants, either randomly selected, or chosen using an IBL model (IBL or Random). Finally, we compared the type of feedback given to participants between positive and negative point feedback and a natural language conversation with an GPT-4o chat-bot (Points or Written).

This experiment included 10 pre-training trials without feedback, 40 training trials with feedback, and 10 post-training trials without feedback. During all trials, participants made judgments of emails as phishing or ham and indicated their confidence in their judgment as well as which action out of 6 possibilities they would select after receiving the email. We recruited 433 participants online through the Amazon Mechanical Turk (AMT) platform. Participants (150 Female, 280 Male, 3 Non-binary) had an average age of 40.3 with a standard deviation of 11.02 years. Participants were compensated with a base payment of $3-5 with the potential to earn up to a $12-15 bonus payment depending on performance and the length of the experiment. This experiment was approved by the Carnegie Mellon University Institutional Review Board, and the study was pre-registered on OSF.

The second component of this dataset is the emails shown to participants, which were either written by human cybersecurity experts, a GPT-4 model working alone, or a combination of human and GPT-4 model work. 360 base emails written by human experts were used to form three additional versions of these base emails. These alternative versions included a 'human-written gpt4-styled' version that used the email body written by human experts, the 'gpt4-written and gpt4-styled' version that was fully rewritten by GPT-4, and the 'gpt4-written plaintext-styled' version that stripped the HTML and CSS styling applied by the GPT-4 model. These emails as well as the original prompts to generate them are included in dataset on OSF.

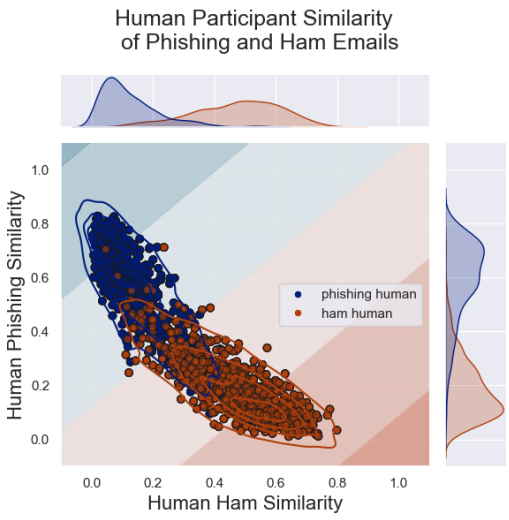

Figure 1: Human participant similarity measure for all 1440 phishing (blue) and ham (orange) emails. Shaded region is a logistic regression.

4 Methods of Measuring Similarity

4.1 Human Subjective Similarity

LLM embeddings have been suggested as a method of measuring human similarity judgements (Bhatia and Aka, 2022), while also capturing the wide range of individuals similarity measures. Additionally, comparisons of LLM behavior have also demonstrated human-like variability (Bhatia, 2024), suggesting these embeddings could be useful for capturing the variety of human similarity judgements. Cognitive models that rely on representations of information from GAI models have been shown to adequately account for the wide range of human behavior (Mitsopoulos et al., 2023).

However, for these methods to function properly there must be a connection between the way that similarity is measured in humans and GAI models. Previous applications in applying visual GAI models onto representing decision-making tasks in humans relied on the close connection to these model representations and human representations (Higgins et al., 2016, 2021). For this reason, we devised a metric of human subjective similarity that takes into account the confidence of document categorization as well as the time it takes participants to categorize documents.

To determine the human subjective similarity measure, we use the category of human participant annotations, their annotation confidence, and the speed of their annotation. For accuracy and confidence, a higher value in our human subjective similarity metric signifies that participants were more likely to categorize an emails as being a member of that group, and more confident in their categorization. For reaction time, a lower value indicates that the document is more immediately obviously a member of a group and thus has a higher similarity to other members of that group. The result is a value that is difficult for a standard similarity metric to account for, as the annotations made in this dataset occurred in a learning setting where earlier trials had less accuracy, which also impacted reaction time and confidence.

The reaction time and confidence weighted subjective similarity of an email x is given by multiplying the probability of a human participant categorizing that email as category c giving $cs(x|c) = p(c|x)r(c|x)c(c|x)$. where $p(c|x)$ is the probability of categorization, $r(c|x)$ is the reaction time normalized to between 0 and 1, and $c(c|x)$ is the confidence additionally normalized to between 0 and 1. The soft-max of this $cs(x|c)$ value is the resulting similarity metric, with the equation shown in the supplementary materials[2].

$$HS(x, x') = \frac{cs(x|c)cs(x'|c)}{\sum_{c' \in C} cs(x|c) \sum_{c' \in C} cs(x'|c')} \quad (5)$$

Figure 1 shows the average human similarity measures for each of the 1440 emails in the dataset. The human ham and human phishing similarities are calculated according to Equation 5 by averaging the accuracy, reaction time, and confidence across all participants in the dataset. It is also possible to calculate this subjective similarity for an individual only using the documents that subject categorized. We next compare similarity measures in their ability to capture individual human subjective similarity.

4.2 Semantic Similarity

One method of measuring the similarity of documents is to employ semantic information contained in documents and compare the similarities and differences between documents in terms of these semantic categories. This has been done in the past in applications such as topic modeling (Řehůřek and Sojka, 2010), document annotation (Pech et al., 2017), and calculating document similarity (Qurashi et al., 2020).

[2] https://osf.io/wbg3r/

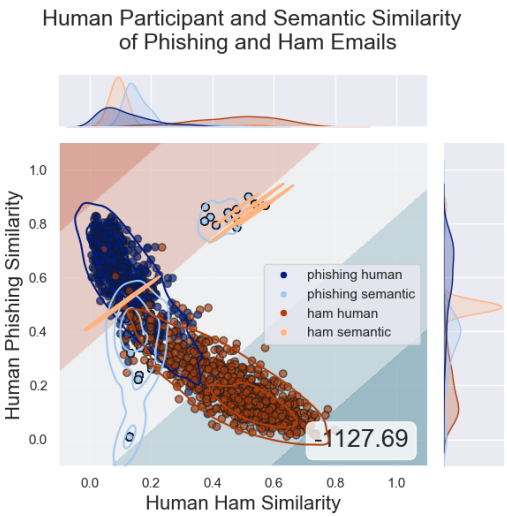

Figure 2: Semantic and human participant similarity for phishing (blue) and ham (orange) emails. Shaded region is a logistic regression. The Kernel Density Estimate log probability score between each distribution is shown on the bottom right, higher is better.

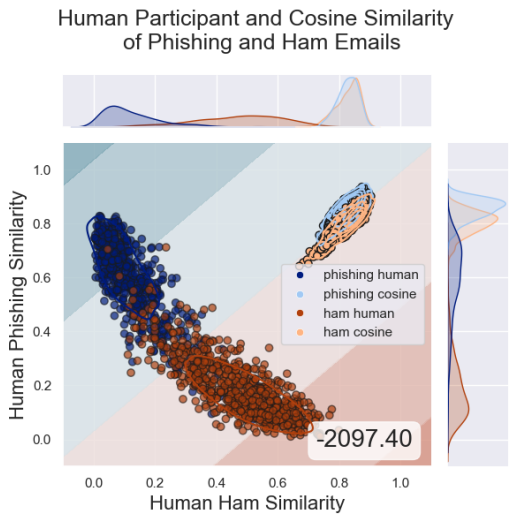

Figure 3: Cosine and human participant similarity for phishing (blue) and ham (orange) emails. Shaded region is a logistic regression. The Kernel Density Estimate log probability score between each distribution is shown on the bottom right, higher is better.

In this dataset, semantic similarity can be calculated using the categorizations of email features that were originally made by the cybersecurity experts who created the base email dataset. These features are Link Mismatch, Offer, Urgent, Subject Suspicious, Request Credentials, and Sender Mismatch. Figure 2 plots these semantic similarity measures for each of the 1440 emails in our dataset, and compares the distribution of these similarities to our human subjective similarity metric.

These semantic similarity metrics are close to human similarity for phishing emails (blue), but highly diverge from the similarity scores of ham emails (orange). This results in a low Kernel Density Estimate log probability score (-1127.69) between the two distributions compared to the semantic similarity metric. This metric compares the likelihood that the data-points in the human similarity metric distribution would have come from the semantic similarity distribution, summing all log probabilities. This low score is due to the fact that the majority of ham emails are very sparse for all of the six semantic categories previously mentioned.

4.3 Cosine Similarity

Cosine similarity is the most commonly used metric of similarity of word and document embeddings, with many applications from classification (Park et al., 2020), recommendation systems (Khatter et al., 2021), educational tutorial systems (Wu et al., 2023), question answering (Aithal et al., 2021), and more (Patil et al., 2023). However, there are limitations to using cosine similarity such as in documents with high-frequency words (Zhou et al., 2022), and the presence of false information (Borges et al., 2019), both of which are concerns for phishing email education.

The cosine similarity metric is calculated using an embedding of size 3072 formed by the 'text-embedding-3-large' model, accessed through the OpenAI API, these document embeddings are additionally included in our presented dataset. The cosine similarity of each email embedding is compared to the mean embedding of that category and shown in Figure 3, and compared to our metric of human subjective similarity. From this, we can see that on average the embeddings are calculated as being significantly more similar to each other compared to the subjective similarities of human participants. This results in a lower Kernel Density Estimate log probability score (-2097.40) between the two distributions compared to the semantic similarity metric.

4.4 Weighted Cosine Similarity

Distance weighted cosine similarity is a common method employed in utilizing embeddings (Li and Han, 2013), which has been applied onto measuring similarity of online instruction in educational settings (Lahitani et al., 2016), as well as several cy-

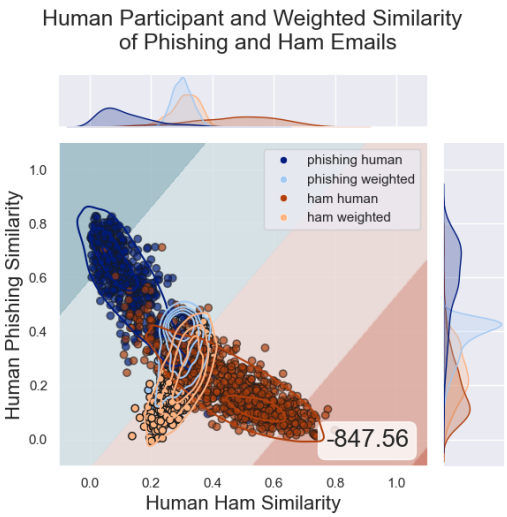

Figure 4: Cosine and human participant similarity for phishing (blue) and ham (orange) emails. Shaded region is a logistic regression. The Kernel Density Estimate log probability score between each distribution is shown on the bottom right, higher is better.

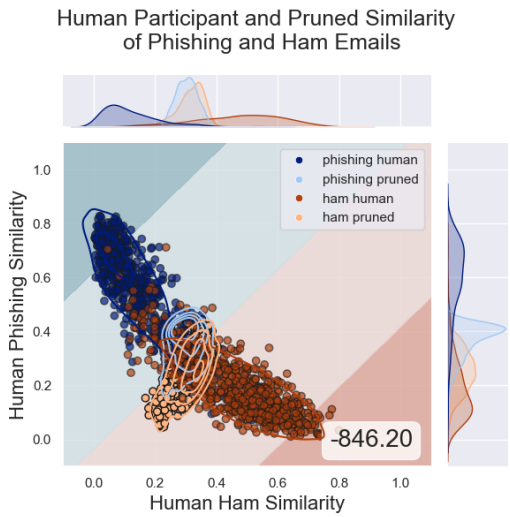

Figure 5: Pruned cosine and human participant similarity for phishing (blue) and ham (orange) emails. Shaded region is a logistic regression. The Kernel Density Estimate log probability score between each distribution is shown on the bottom right, higher is better.

bersecurity specific applications like ransomware detection (Moussaileb et al., 2021), and inside attacker detection (Khan et al., 2019). In this work, we employ weighted cosine similarities of embeddings formed from emails categorized as being either ham or phishing, and compare it to human subjective similarity judgements. This weighting is done by learning a weight transformation of size 3072, the same as the embedding size, which is applied onto the embedding prior to calculating the similarity. The results of this weighting are shown in Figure 4, which compares the average human participant subjective similarity and the weighted cosine similarity of email embeddings.

The KDE log probability score between weighted cosine similarities of phishing and ham emails compared to human subjective similarity has increased to -847.56 from the unweighted KDE score of -2097.40, surpassing the semantic similarity score at -1127.69. These improved similarity metrics indicate that weighting cosine similarity based on data from a large dataset of human participants can result in a metric that more accurately reflects the average of human subjects' subjective similarity metrics.

4.5 Pruning Document Embeddings

Another method of comparison documents is embedding pruning, where embeddings are reduced in size based on feedback from human categorizations to better account for their subjective similarity (Manrique et al., 2023). These approaches function by reducing the number of embedding values that are used in comparison, and are similar to the weighting method except with 0 or 1 values. We structured our embedding pruning method to select only the top 500 embedding values, representing just under 20% of the size of the embedding, as was done in (Manrique et al., 2023). These top predictive embedding values are retained, while all other values are masked to 0. After this, cosine similarity can be calculated with the standard approach, resulting in the similarity shown in Figure 6. Compared to the weighted cosine similarity method, the pruned cosine similarity has roughly the same KDE log probability score.

5 Ensemble Similarity

The final comparison method is based on using an ensemble of each of the previous similarity metrics, weighted to maximize the similarity to the average of the human subjective similarity metrics. This approach has been applied to document matching for patent documents (Yu et al., 2024), which requires the similarity of document embeddings be calculated to determine a match. This ensemble approach has the highest KDE log probability score of any individual method by itself, at a value of -812.23. Looking at the KDE distributions above and to the right of the scatter plot in 6 demonstrates

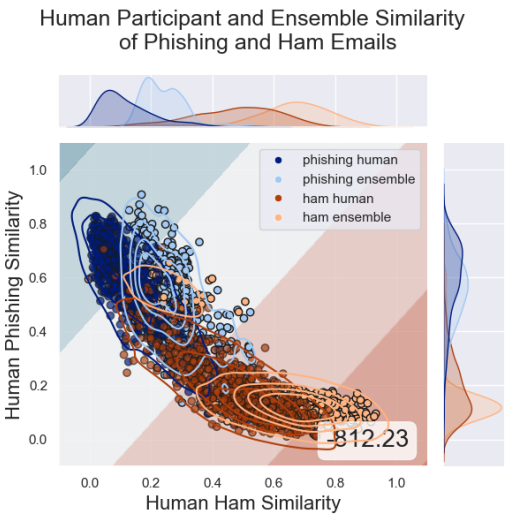

Figure 6: Ensemble and human participant similarity for phishing (blue) and ham (orange) emails. Shaded region is a logistic regression. The Kernel Density Estimate log probability score between each distribution is shown on the bottom right, higher is better.

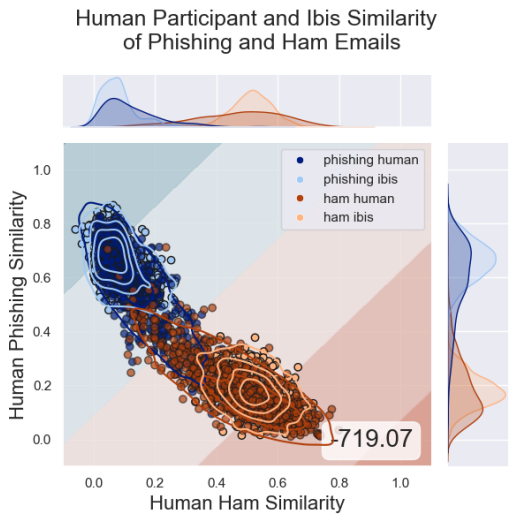

Figure 7: IBIS and human participant similarity for phishing (blue) and ham (orange) emails. Shaded region is a logistic regression. The Kernel Density Estimate log probability score between each distribution is shown on the bottom right, higher is better.

the high similarity of the ensemble similarity metric (light blue and light orange) and the human participant similarity metric (blue and orange). While this method is effective at resulting in a similarity metric that closely matches the average over all participants, it still does not fit as well to individual participants, as will be shown in our proposed model.

6 Instance-Based Individualized Similarity (IBIS)

To determine an individual participant's metric of similarity, we employ an IBL model that is serving as a digital twin of the participant. The result in an Instance-Based Individualized Similarity (IBIS) metric. The benefits of IBIS are in the ability to predict human judgements on unseen documents or feedback from recommendations, and enhance measurements of subjective similarity. Importantly, these predictions of human behavior are not merely relying on a separate machine learning based technique, but rather a cognitive model that is inspired by the human cognitive mechanisms underlying decision making and thus able to account for natural biases and constraints in humans.

Predictions of Instance-Bases Individual Similarity are done using an IBL model that is currently serving as a digital twin with the same experience as an individual participant. Using this we determine the value that the IBL model assigns to predicting a category c as $V_k(c|x)$, or the value the IBL model assigns to choosing option c as the category of document x. Then, we can divide this value by the same categorization value assigned to each alternative categorization of the same document. This results in the IBIS metric which can be calculated after each decision is made by a participant, pseudo-code for the IBIS algorithm, The code-base for the IBIS method including all comparison methods, data, and scripts to generate similarity measures and figures is made available[3].

7 Case Study of IBIS: Individuals in Phishing Email Education Dataset

Previous comparisons of similarity metrics and human participant behavior compared the average of human performance. To highlight the benefits of the IBIS method, we replicate these calculations with one individual from the experiment. Here, the individual similarity of phishing and ham emails is based only on a single individuals categorization, confidence, and reaction time in their judgement. These graphs are shown for illustration in Figure 8, with the average accuracy of logistic regression of similarity metrics predicting individual participant similarity metrics reported in table 1.

The KDE score of the similarities for the pruned cosine method is -30.76, and for the ensemble method it is -27.82. Note that these scores are much

[3] github.com/TylerJamesMalloy/cognitive-similarity

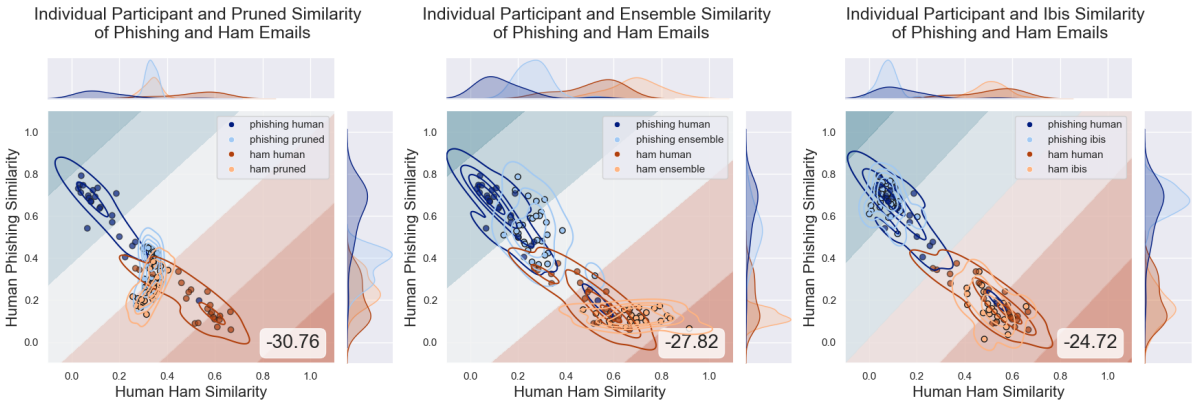

Figure 8: Top performing similarity metrics and individual participant similarity for phishing and ham emails. Shaded region is a logistic regression. The lower value is the individual KDE score

lower than the entire dataset scores since they are calculated using only the emails observed by the participant. Meanwhile, the IBIS metric gives a KDE score of -24.72. From this we can see that the IBIS method effectively learns the similarity measures of individual participants. These results are used for illustrative purposes, and the averages across all participants for regression accuracy, as well as the DKE score for individuals, is presented in Table 1.

An important aspect of individual similarity comparisons of the IBIS method is that it can compare emails that were not originally presented to an individual, meaning there are more embedding similarities used in the logistic regression and KDE score calculation. This comparison demonstrates the benefits of using a cognitively inspired method of modeling human participant decisions making that takes into account biases and cognitive constraints.

The results is a prediction of behavior that can accurately fill in the gaps of unseen elements of the dataset that have not been observed by a participant. This method more accurately predicts the subjective similarity of participants. Importantly, this is done while initially limiting the cognitive model to observing a single decision made by these participants, and increasing this data as the participant makes more decisions. This is important for the functioning of the IBL model as using too many instances in memory can slow compute performance.

The final comparison shown in the right most columns of Table 1 shows the percent accuracy in using the previously described logistic regressions, shown on all figure results, in predicting the categorization of participants based on the similarity metric applied onto the emails they observed. This regression has the potential to predict the annotations of individuals, similarly to the IBL model. Comparing these measures shows that the best performance comes from the IBIS metric when predicting participant annotations.

8 Discussion

Many applications of LLMs are interested in tailoring use cases to individuals, even when little information is known about that individual. While many approaches of individualization exist but have typically relied on advanced machine learning techniques. The method proposed in this work is relatively simple from a mathematical perspective, though there is a strength in its reliance on theories of cognition that underlie human learning and decision making. The result is a simple to understand and easy to implement method of calculating similarities of unseen documents using a cognitive model, which can augment datasets that contain only a small number of decisions.

The general method described here, of augmenting subjective similarity metrics with predicted decisions from a cognitive model, could be applied onto various other scenarios. This includes settings that leverage representations formed of visual information such as β-Variational Autoencoders (Higgins et al., 2016), which have been related to biological representation formation (Higgins et al., 2021). Overall, we believe that this method is useful for any application where the experience of end-users impacts future decisions.

For instance, in visual learning settings VAEs have been integrated with cognitive models to predict human utility learning of abstract visual in-

Similarity Metric	KDE Score Average Participants	KDE Score Individuals	Regression Accuracy
Semantic Similarity (Qurashi et al., 2020)	-1127.69	-37.69±1.19	0.46±0.11
Cosine Similarity (Park et al., 2020)	-2097.40	-47.26±2.27	0.52±0.10
Embedding Weighting (Onan, 2021)	-847.56	-29.28±2.32	0.86±0.14
Embedding Pruning (Manrique et al., 2023)	-846.20	-30.39±2.76	0.86±0.04
Ensemble Similarity (Yu et al., 2024)	-812.23	-28.64±3.28	0.89±0.12
IBIS (proposed)	**-719.07**	**-23.17±3.29**	**0.93±0.04**

Table 1: Comparison of the six previously described methods in their similarity to human behavior. Similarity to average participants is performed across the entire dataset of human judgements (see Figures 1-6). Similarity to individuals and regression accuracy are both done for each individual participant (see Figure 7). For all values higher is better. Reported values are means of all participants measured individually ± standard deviations.

formation (Malloy and Sims, 2024). Other integrations of Generative AI into cognitive models includes use of LLMs as a knowledge repositories within cognitive models (Kirk et al., 2023). In particular, ConceptNet (Speer et al., 2017) has previously been integrated into a cognitive model for question answering (Huet et al., 2021). Future research should investigate how additional uses of LLMs in integrations of cognitive models can aid in educational settings.

Overall, the results in this work demonstrate the usefulness of cognitive models in serving as digital twins to human participants. Leveraging these models and integrating their results into Large Language Model techniques has the potential to make measurements from these models more cognitively grounded. While there are existing methods of incorporating human behavior through the use of large datasets collected from many participants, these do not necessarily account for biases and constraints. The method proposed in this work takes these features of human learning and decision making into account in developing a similarity metric.

9 Limitations

The semantic similarity metric suffered from the sparsity of semantic categories in ham emails, additional annotations could raise the performance of this metric and can be explored in future work. However, this ensemble method was partially responsible for the high KDE score of the ensemble method, as it allowed for an integration of both semantic information and embedding similarity. Our IBIS method still outperformed the ensemble method suggesting that this ensemble alone does not address the issues of alternative methods.

One limitation inherent in IBL cognitive models is the time requirements to compare the current instance to all instances in memory. This may make the proposed model unsuitable for applications that rely on large datasets of individual behavior. However, methods in instance compression exist for IBL models (Nguyen et al., 2023). In this setting, we were able to predict individual participant's decisions fast enough that this was unnecessary.

The specific application we investigated is somewhat unique in that it is based on training human participants to make categorization judgements of textual information of one of two categories. Additionally, the task of annotating whether an email is phishing or ham relies heavily on a small number of features within the email. Namely, if an email contains a link that redirects to a nefarious website, or requests personal information, then it should be labelled as phishing. While students rely on many queues to make their judgements, the annotation is in reality simple. Future work in the area of learning subjective similarity metrics should expand into more complex domains.

10 Acknowledgement

This research was sponsored by the Army Research Office and accomplished under Australia-US MURI Grant Number W911NF-20-S-000, and the AI Research Institutes Program funded by the National Science Foundation under AI Institute for Societal Decision Making (AI-SDM), Award No. 2229881. Compute resources and GPT model credits were provided by the Microsoft Accelerate Foundation Models Research Program grant "Personalized Education with Foundation Models via Cognitive Modeling".

References

Palvi Aggarwal, Omkar Thakoor, Shahin Jabbari, Edward A Cranford, Christian Lebiere, Milind Tambe, and Cleotilde Gonzalez. 2022. Designing effective masking strategies for cyberdefense through human experimentation and cognitive models. *Computers & Security*, 117:102671.

Shivani G Aithal, Abishek B Rao, and Sanjay Singh. 2021. Automatic question-answer pairs generation and question similarity mechanism in question answering system. *Applied Intelligence*, pages 1–14.

Sudeep Bhatia. 2024. Exploring variability in risk taking with large language models. *Journal of Experimental Psychology: General*.

Sudeep Bhatia and Ada Aka. 2022. Cognitive modeling with representations from large-scale digital data. *Current Directions in Psychological Science*, 31(3):207–214.

Luís Borges, Bruno Martins, and Pável Calado. 2019. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–26.

Peter Brown and Yaoqi Zhou. 2019. Large expert-curated database for benchmarking document similarity detection in biomedical literature search. *Database*, 2019:baz085.

Carolina Carreira, McKenna McCall, and Lorrie Faith Cranor. 2024. How to explain trusted execution environments (tees)? In *USENIX Symposium on Usable Privacy and Security (SOUPS)*.

Stephen L Chew and William J Cerbin. 2021. The cognitive challenges of effective teaching. *The Journal of Economic Education*, 52(1):17–40.

Edward A Cranford, Cleotilde Gonzalez, Palvi Aggarwal, Sarah Cooney, Milind Tambe, and Christian Lebiere. 2020. Toward personalized deceptive signaling for cyber defense using cognitive models. *Topics in Cognitive Science*, 12(3):992–1011.

Edward A Cranford, Christian Lebiere, Prashanth Rajivan, Palvi Aggarwal, and Cleotilde Gonzalez. 2019. Modeling cognitive dynamics in end-user response to phishing emails. *Proceedings of the 17th ICCM*.

Achraf Gazdar and Lotfi Hidri. 2020. A new similarity measure for collaborative filtering based recommender systems. *Knowledge-Based Systems*, 188:105058.

Cleotilde Gonzalez and Varun Dutt. 2011. Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological review*, 118(4):523.

Cleotilde Gonzalez, Javier F Lerch, and Christian Lebiere. 2003. Instance-based learning in dynamic decision making. *Cognitive Science*, 27(4):591–635.

Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. 2021. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):6456.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, pages 1–6.

Armand Huet, Romain Pinquié, Philippe Véron, Antoine Mallet, and Frédéric Segonds. 2021. Cacda: A knowledge graph for a context-aware cognitive design assistant. *Computers in Industry*, 125:103377.

Ahmed Yar Khan, Rabia Latif, Seemab Latif, Shahzaib Tahir, Gohar Batool, and Tanzila Saba. 2019. Malicious insider attack detection in iots using data analytics. *IEEE Access*, 8:11743–11753.

Harsh Khatter, Nishtha Goel, Naina Gupta, and Muskan Gulati. 2021. Movie recommendation system using cosine similarity with sentiment analysis. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 597–603. IEEE.

James R Kirk, Robert E Wray, and John E Laird. 2023. Exploiting language models as a source of knowledge for cognitive agents. *arXiv preprint arXiv:2310.06846*.

Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. 2022. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. 2016. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International conference on cyber and IT service management*, pages 1–6. IEEE.

Tomás Lejarraga, Varun Dutt, and Cleotilde Gonzalez. 2012. Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, 25(2):143–153.

Baoli Li and Liping Han. 2013. Distance weighted cosine similarity measure for text classification. In *Intelligent Data Engineering and Automated Learning–IDEAL 2013: 14th International Conference, IDEAL 2013, Hefei, China, October 20-23, 2013. Proceedings 14*, pages 611–618. Springer.

Tyler Malloy, Maria Ferriera Jose, Fei Fang, and Cleotilde Gonzalez. 2024. Improving online anti-phishing training using cognitive large language models. *Under Review for Computers in Human Behavior*.

Tyler Malloy and Cleotilde Gonzalez. 2024. Applying generative artificial intelligence to cognitive models of decision making. *Frontiers in Psychology*, 15:1387948.

Tyler Malloy and Chris R Sims. 2024. Efficient visual representations for learning and decision making. *Psychological review*.

Natalia Flechas Manrique, Wanqian Bao, Aurelie Herbelot, and Uri Hasson. 2023. Enhancing interpretability using human similarity judgements to prune word embeddings. *arXiv preprint arXiv:2310.10262*.

Konstantinos Mitsopoulos, Rik Bose, Brodie Mather, Archna Bhatia, Kevin Gluck, Bonnie Dorr, Christian Lebiere, and Peter Pirolli. 2023. Psychologically-valid generative agents: A novel approach to agent-based modeling in social sciences. In *Proceedings of the 2023 AAAI Fall Symposium on Integrating Cognitive Architectures and Generative Models*, pages 1–6. AAAI Press.

Routa Moussaileb, Nora Cuppens, Jean-Louis Lanet, and Hélène Le Bouder. 2021. A survey on windows-based ransomware taxonomy and detection mechanisms. *ACM Computing Surveys (CSUR)*, 54(6):1–36.

Shaimaa M Nafea, François Siewe, and Ying He. 2019. A novel algorithm for course learning object recommendation based on student learning styles. In *2019 International Conference on Innovative Trends in Computer Engineering (ITCE)*, pages 192–201. IEEE.

Thuy Ngoc Nguyen and Cleotilde Gonzalez. 2022. Theory of mind from observation in cognitive models and humans. *Topics in Cognitive Science*, 14(4):665–686.

Thuy Ngoc Nguyen, Duy Nhat Phan, and Cleotilde Gonzalez. 2023. Speedyibl: A comprehensive, precise, and fast implementation of instance-based learning theory. *Behavior Research Methods*, 55(4):1734–1757.

Aytuğ Onan. 2021. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and computation: Practice and experience*, 33(23):e5909.

Saurabh Pal, Pijush Kanti Dutta Pramanik, and Prasenjit Choudhury. 2024. Aggregated relative similarity (ars): a novel similarity measure for improved personalised learning recommendation using hybrid filtering approach. *Multimedia Tools and Applications*, pages 1–48.

Kwangil Park, June Seok Hong, and Wooju Kim. 2020. A methodology combining cosine similarity with classifier for text classification. *Applied Artificial Intelligence*, 34(5):396–411.

Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. 2023. A survey of text representation and embedding techniques in nlp. *IEEE Access*.

Fernando Pech, Alicia Martinez, Hugo Estrada, and Yasmin Hernandez. 2017. Semantic annotation of unstructured documents using concepts similarity. *Scientific Programming*, 2017(1):7831897.

Abdul Wahab Qurashi, Violeta Holmes, and Anju P Johnson. 2020. Document processing: Methods for semantic text similarity analysis. In *2020 international conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6. IEEE.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Mansoore Shojaei and Hassan Saneifar. 2021. Mfsr: A novel multi-level fuzzy similarity measure for recommender systems. *Expert Systems with Applications*, 177:114969.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, pages 1–6.

Xuansheng Wu, Xinyu He, Tianming Liu, Ninghao Liu, and Xiaoming Zhai. 2023. Matching exemplar as next sentence prediction (mensp): Zero-shot prompt learning for automatic scoring in science education. In *International conference on artificial intelligence in education*, pages 401–413. Springer.

Peipei Xia, Li Zhang, and Fanzhang Li. 2015. Learning similarity with cosine similarity ensemble. *Information sciences*, 307:39–52.

Liqiang Yu, Bo Liu, Qunwei Lin, Xinyu Zhao, and Chang Che. 2024. Semantic similarity matching for patent documents using ensemble bert-related model and novel text processing method. *arXiv preprint arXiv:2401.06782*.

Haiqin Zhang, Zheng Chen, Wei-ying Ma, and Qingsheng Cai. 2003. A study for document summarization based on personal annotation. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 41–48.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. *arXiv preprint arXiv:2205.05092*.