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

Anonymous EMNLP submission

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 (Ansari et al., 2000) and educational tools (Nafea et al., 2019) typically do not have personalized measurements of subjective concepts, either for themselves or the people that are using these systems, potentially hindering their efficacy.

The specific use case we are interested in is a learning setting where students are categorizing documents and receiving feedback of the accuracy of their categorization. In this work, 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.

When these systems do incorporate data from human judgements to determine the subjective similarity of, 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 can be effective from a machine learning perspective, since more data can mean improved performance for the general public. But in terms of providing an individualized experience to students in educational settings or end-users in recommendation settings, this type of data aggregation approach leaves something to be desired.

When methods do attempt to account for individual measures of similarity, they typically employ machine learning based methods (Shojaei and Saneifar, 2021). While these approaches can be beneficial in some use cases, they are not grounded by the biases and constraints inherent in human learning in a way that is afforded through cognitive modeling.

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 textual categorization and educational 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 behavior and opinions on a wider set of stimuli. This cognitive model incorporates LLM embeddings into its prediction of human behavior, allowing for flexible and efficient connections between cognitive models and LLMs.

Other than this method, the main contribution of this work is in presenting a dataset of human participant judgements in an educational setting learning to correctly categorize emails as being either safe 'ham' or dangerous 'phishing'. This is a nuanced and difficult task that could potentially be based on multiple different subjective classifications such as the level of urgency in a text, whether it has a suspicious tone, or whether it is making an offer that seems too good to be true. All of these features are relevant to determining if an email is genuine, and the way that individuals perceive an email as having these features can be highly individualistic.

Alongside data from these human judgements we present a dataset of emails written by human cybersecurity experts, as well as emails generated by GPT-4 while relying on various levels of information from human prompt engineers. The final part of this dataset is a set of conversations between human participants and a GPT-4o model providing feedback to students. These conversations and the resulting educational improvement of students can provide useful insight into the prompting of LLMs for educational settings.

In total this dataset represents 39230 human judgements from 430 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.

2 Phishing Email Categorization Dataset

One of the contributions of this work is the presentation of a dataset of human judgements categorizing emails as being either phishing or ham. These emails were created from various sources, including human cybersecurity experts, GPT-4 generation, as well as humans working with GPT-4 in collaboration. Data from human participants categorizing these emails as being either ham or phishing in an educational setting was made available online at [1]. This dataset includes 39230 categorization judgements from 384 human participants of 1440 possible emails.

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 the presented dataset.

The final component of this dataset is a set of conversations between human participants and a GPT-4o model prompted to teach the participant to identify phishing emails. In this experiment three out of the eight experimental conditions involved human participants discussing the emails that they were categorizing with a GPT-4o model. This model was prompted to serve as an educational tool and varied in the type of information that was included in these prompts across experimental conditions. These teacher-student conversations consist of 20487 messages sent between human participants and the GPT-4o model.

3 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). This learning theory describes the mathematical foundation of cognitive mechanisms that underlie human decision making in dynamic environments, such as learning tasks. Cognitive models that rely on the mathematical framework of IBLT are called Instance-Based Learning models, which are used to define the features relevant for decision making tasks, and to predict the mechanisms of dynamic decision making based on these features.

One of the benefits of employing IBL models over alternatives is that they take into account the past experiences of participants and the impact of limitations like memory size and decay that can bias decision making. IBL models have been applied onto predicting human behavior in dynamic decision making tasks, including repeated binary choice tasks (Gonzalez and Dutt, 2011; Lejarraga et al., 2012), theory of mind applications (Nguyen

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

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). The following sections outline the mathematical foundation of IBL models, and gives attention to the method of integrating these concepts into predictions of subjective similarity of categories.

3.1 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 $\mathcal{T}(i)$. IBL models predict the value of options in decision-making tasks by selecting the action that maximizes the value function. 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.

3.2 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:

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

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 $\tau = \sigma\sqrt{2}$.

3.3 Blended Value

The blended value of an option k is calculated 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:

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

The blended value of different options is key to predicting the subjective similarity of human participants in a way that is both individualized, and takes into account the experience of students in educational settings. This is done in our proposed individualized metric of subjective similarity by comparing the blended value of an individual that is engaged in a task to categorize emails as being either phishing or spam.

4 Methods of Measuring Similarity

4.1 Human Similarity Measures

We can use the accuracy of human participant categorizations and the confidence that participants selected to their judgement to plot for each email their level of similarity to phishing and ham emails. For both of these metrics, a higher value signifies that participants were more likely to categorize an emails as being a member of that group with a high confidence. These results are graphed in Figure 1, which plots the phishing and ham similarity of each email based on the average of human performance.

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. This gives the subjective similarity as:

$$HS(x, x') = \frac{cs(x|c)cs(x'|c)}{\sum_{c' \in C} cs(x|c) \sum_{c' \in C} cs(x'|c')} \quad (4)$$

This metric of subjective similarity depends on primarily the categorization of emails from human

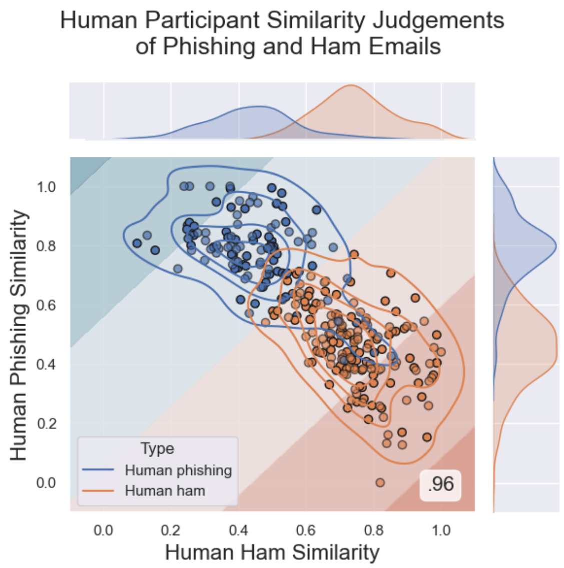

Figure 1: Average Human Similarity measure for phishing (blue) and ham (orange) emails based on the categorization, confidence, and response latency of participant responses. Shaded region represents a logistic regression classifier trained on 100 train-test splits of size 50% with the accuracy shown in the lower right.

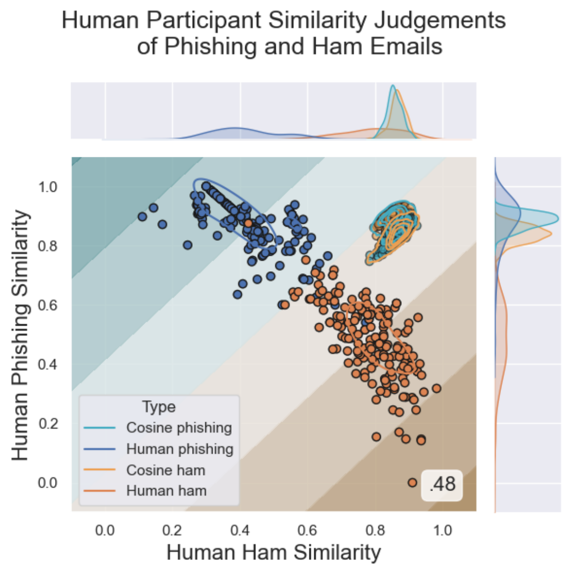

Figure 2: Average cosine and human participant similarity for phishing (light blue) and ham (light orange) emails. Shaded region represents a logistic regression classifier trained on 100 train-test splits of size 50% with the accuracy shown in the lower right.

participants, but additionally takes into account their confidence, with higher confidences of phishing categorizations indicating an emails is more similar to the phishing email category. Additionally, this metric takes into account the reaction time of human participants in making their judgement, with faster judgements additionally indicating that a category is more similar to members of that category average. The goal of the IBIS method is to reflect this type of subjective similarity, and is compared to several alternative measures of similarity described in the following sections.

4.2 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.

A simple way to apply cosine similarity onto the task of categorizing an email as being either phishing or ham is to collect a large number of labelled emails and compute the average of the embeddings of these labelled emails. Once this embedding average is collected, we can measure the cosine distance of any given email embedding and the average of both categories.

$$\begin{aligned} \mathrm{CS}(x, x') &= \frac{x^T x'}{||x||||x'||} \\ &= \frac{x^T x'}{\sqrt{x^T x}\sqrt{x'^T x'}} \end{aligned} \tag{5}$$

The cosine similarity of each email embedding to the mean embedding of that category is shown in Figure , and compared to our metric of subjective similarity that is dependent on human participant categorization, confidence, and reaction time. 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.

Comparing the accuracy of using the cosine similarity metrics, we can see that the logistic regression of predicting the human subjective similarity has now decreased in accuracy to 0.48, from the previous accuracy of 0.96 when forming a logistic regression of human participant similarity judgements alone. This significant decrease is due to the gap between cosine similarity and the subjective

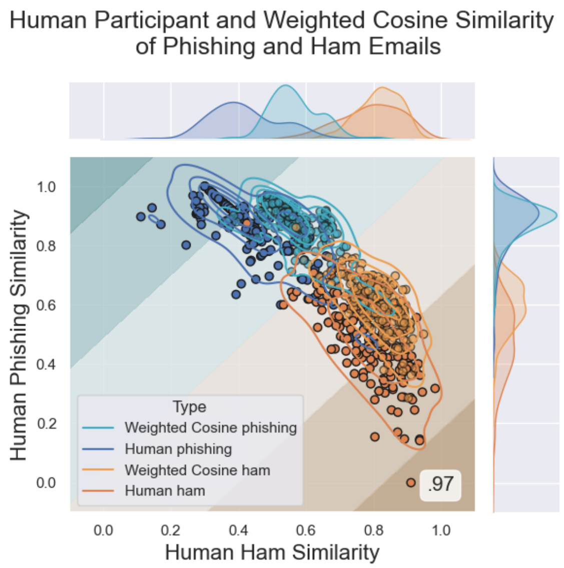

Figure 3: Average weighted cosine and human participant similarity for phishing (light blue) and ham (light orange) emails. Shaded region represents a logistic regression classifier trained on 100 train-test splits of size 50% with the accuracy shown in the lower right.

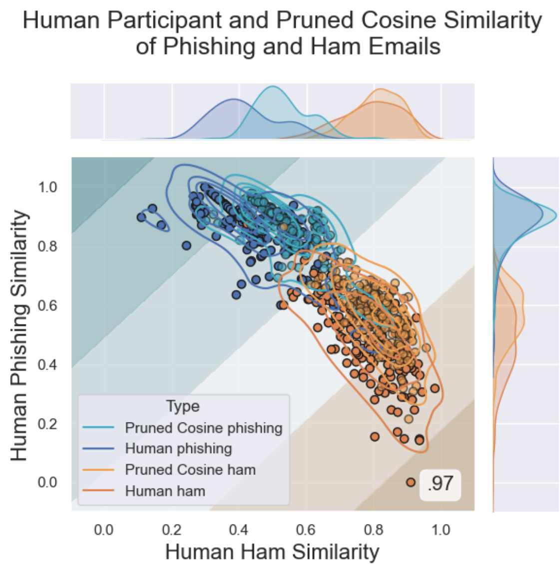

Figure 4: Average pruned cosine and human participant similarity for phishing (light blue) and ham (loght orange) emails. Shaded region represents a logistic regression classifier trained on 100 train-test splits of size 50% with the accuracy shown in the lower right.

similarity of human participants. One solution to this is to deferentially weight the indices of embedding values, which is explored next.

4.3 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 cybersecurity specific applications like ransomeware 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 can be done by defining the weighted cosine similarity of an email embedding as:

$$\begin{aligned} \mathrm{CS}_w(x, x', W) &= \frac{(Wx)^T(Wx')}{||Wx||||Wx'||} \\ &= \frac{x^T W^T W x'}{\sqrt{x^T W^T W x}\sqrt{x'^T W^T W x'}} \end{aligned} \tag{6}$$

From this definition of the weighted cosine similarity, it is relatively simple to construct the embedding weight matrix A in a way that minimizes the mean squared error of the distance between weighted cosine embeddings and the subjective similarity metrics of participants. This allows for the classification of emails in a way that reflects the confidence and categorization of human participants in an educational setting. The results of this weighting are shown in Figure 3, which compares the average human participant subjective similarity and the weighted cosine similarity of email embeddings.

The accuracy of the logistic regression fit to weighted cosine similarities of phishing and ham emails when predicting human subjective similarity has increased to 0.97 from the unweighted accuracy of 0.48. 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.

4.4 Pruning Document Embeddings

The final method of comparison for developing individualized metrics of similarity 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). This method was originally designed for word embeddings with a larger number of categories that are more varied than our application. We adjusted this approach to apply it onto reflecting human categorization of emails into only two related categories of phishing and ham emails.

After making these adjustments to the embedding pruning method the result is a similarity metric calculated by ranking embedding value by how well it predicts the different human categorization performance, and selecting 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 5.

5 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.

$$IBIS(x, x') = \frac{V_k(c|x)V_k(c|x')}{\sum_{c' \in C} V_k(c'|x) \sum_{c' \in C} V_k(c'|x')} \tag{7}$$

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, as shown in the pseudo-code in Algorithm 1.

Input: default utility u_0, a memory dictionary $\mathcal{M} = \{\}$, global counter $t = 1$, step limit L. Dataset of stimuli D

repeat
- Initialize a counter (i.e., step) $l = 0$ and observe state s_l
- **while** *s_l is not terminal and $l < L$* **do**
  - **Execution Loop**
    - **Exploration Loop** $k \in K$ **do**
      - Compute $A_i(t)$ by Eq: (1)
      - Compute $P_i(t)$ by Eq: (2)
      - Compute $V_k(t)$ by Eq: (4)
    - **end**
    - Update similarity by Eq: (7) using each data point in D
    - Predict student action a by $k_l \in \arg\max_{k \in K} V_k(t)$
  - **end**
  - Observe student action a, observe s_{l+1}, and student feedback outcome u_{l+1}
  - Store t instance in $\mathcal{M}$
- **end**

until *task stopping condition*

Algorithm 1: Pseudo Code of Instance-Based Learning Cosine Similarity Update

6 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 with the average accuracy of logistic regression of similarity metrics predicting individual participant similarity metrics reported in table 1.

While the previous comparisons of embedding similarity metrics were all reasonably reflective of the average of all human participants across the entire dataset, they do not necessarily correspond to individual participants as closely. To demonstrate this, we plot 5 randomly selected participants' individual metrics of similarity for the limited emails they observed in Figure 5. Here, the individual

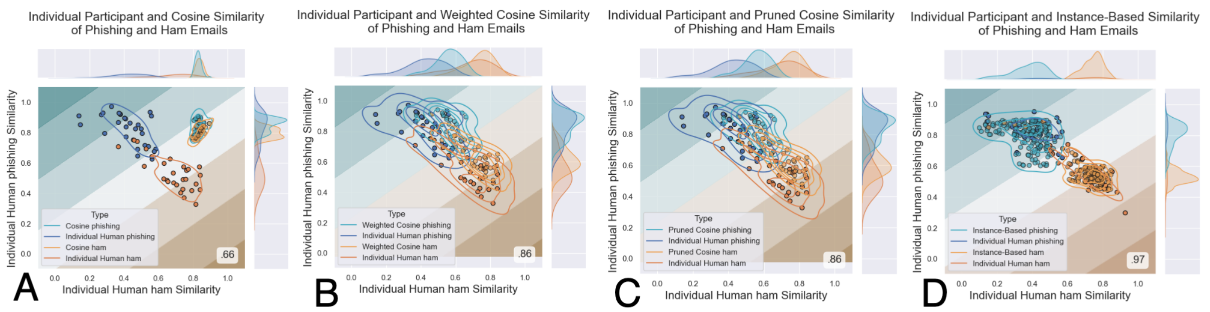

Figure 5: A comparison of the four similarity metrics under comparison using data from a single individual participant. These are shown for illustrative purposes with averages for each participant listed in Table 1 A: Cosine similarity compared to human subjective similarity. B: Weighted cosine similarity compared to human subjective similarity. C: Pruned cosine similarity compared to human subjective similarity. D: IBIS similarity compared to human subjective similarity.

similarity of phishing and ham emails is based only on a single individuals categorization, confidence, and reaction time in their judgement. From this we can see that there is a large discrepancy between the aggregated weighted cosine similarity and each of the four individual participants.

The accuracy of the logistic regression of the embeddings for the vanilla cosine similarity for this example participant is 0.66. The same value for both the weighted cosine and pruned cosine method for this participant is 0.86. Meanwhile, the IBIS metric gives an accuracy of the logistic regression of 0.96. This is approaching the original accuracy of the two best performing cosine similarity metrics (weighting and pruning) when using the entire dataset of human participant performance.

An important difference between these four methods is that only the IBIS method can compare emails that were not originally presented to an individual, meaning there are more embedding similarities used in the logistic regression. Note the smaller number of similarities performed in the Cosine, Weighted Cosine, and Pruned Cosine conditions in the first three columns of Figure 5, which is due to the limited number of emails shown to each individual participant. Meanwhile, the IBIS method has a larger sample of emails to draw from since it makes predictions of individual participant behavior on emails that they were never presented with.

This comparison demonstrates the clear 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 as measured by categorization, confidence, and reaction time. 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.

7 Predicting Human Categorization

To evaluate the usefulness of these previously listed metrics of semantic similarity, we employ another IBL model to make predictions of participant categorizations of phishing emails as being either safe or dangerous. This is a highly complex task that involves making judgements about subjective qualities like suspicousness, urgency, or plausibility, as well as objective qualities like whether the email sender matches one listed in the body or whether a link url is mismatched from its text.

In the above examples of using different methods to predict the subjective similarity of human participant behavior, the entire dataset of decisions from one individual was used. However, when providing educational feedback the number of data points from each participant begins at 0 and progresses through to the full amount of decisions collected from that participant. This is a significantly more challenging problem, as human participant decisions can be poor at the beginning of educational examples and potentially increase in quality dramatically through educational feedback.

To compare the methods discussed in this paper, as well as our proposed Instance-Based cosine similarity weighting approach, we evaluate the accuracy of logistic regressions as they are formed

Method	Similarity to Average Humans	Similarity to Individual Humans	Human Behavior IBL Prediction Accuracy
Cosine Similarity (Park et al., 2020)	0.48	0.60±0.2	0.80±0.1
Embedding Weighting (Li and Han, 2013)	0.97	0.86±0.1	0.81±0.04
Embedding Pruning (Manrique et al., 2023)	0.97	0.86±0.04	0.82±0.08
IBIS (proposed)	0.97	**0.93±0.04**	**0.87±0.05**

Table 1: Comparison of the three previously described methods in their similarity to human behavior. Similarity to average humans is performed across the entire dataset of human judgements. Similarity to individuals and IBL prediction accuracy are both done for each individual participant. Reported values are means ± standard deviations.

from individual participants behavior. This is done by comparing the regression accuracy in predicting the single next decision made by a human participant while fitting the measure of their subjective similarity from all previous decisions that they have made.

The results from this comparison of the predictive accuracy of a separate IBL model that relies on different metrics of similarity when predicting human performance are shown in Table 1, and indicate that the IBIS method of calculating individualized subjective similarity of participants produces the best similarity metric for an IBL model when predicting human behavior. It is important to note that the IBL model predicting human behavior and the model that are estimating similarity are not the same, as the similarity estimate model needs to rely on a separate similarity metric.

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 have demonstrated success in producing outputs or representing information in an individualized manner, these 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 from a single user.

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. However, we believe that the general method described, of augmenting subjective similarity metrics with predicted decisions from a cognitive model, could be applied onto various other scenarios.

For instance, in visual learning settings Variational Autoencoders have been integrated with cognitive models to predict human utility learning of abstract visual information (Malloy and Sims, 2024). This task involved visual queues with associated utilities taken from a large dataset of hundreds of possible abstract visual images in the form of jars of differently colored marbles. The same method of determining subjective similarity could be applied onto this visual utility learning task.

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 individual biases and constraints. The method proposed in this work takes these features of human learning and decision making into account in developing individualized metrics of similarity.

Limitations

The task presented in this work of predicting 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 true categorization task is in reality simple. Future work in the area of learning subjective similarity metrics should expand into domains with more categories, and more complex and abstract categories.

Ethics Statement

The model proposed in this work, as well as the dataset introduced, involves an educational setting and thus introduces significant ethical concerns. One of the main concerns of the use of LLMs in educational settings is the potential for biases present in LLMs that negatively impact students of a specific ethnic, cultural, or racial background. This potential concern is mitigated in this work because of the specific educational setting, in detecting phishing emails, which are designed by the original cybersecurity experts to be applicable to a wide range of end users. However, the application of this approach outside of the setting used in this work should take care in ensuring that the method of calculating the similarity of educational examples shown to students not be biased. While this is an inherent concern in the use of LLMs in education, our proposed approach of using more individualized metrics of similarity can hopefully reduce the likelihood of LLM biases negatively impacting student education. This is because our proposed model is based on individual past experiences and biases when calculating subjective similarity.

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.

Asim Ansari, Skander Essegaier, and Rajeev Kohli. 2000. Internet recommendation systems.

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.

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*.

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.

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.

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 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*, in press.

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*.

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.

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*.

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

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.

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*.