KnowledgeVIS: Interpreting Language Models by Comparing Fill-in-the-Blank Prompts

Adam Coscia and Alex Endert

Abstract—Recent growth in the popularity of large language models has led to their increased usage for summarizing, predicting, and generating text, making it vital to help researchers and engineers understand how and why they work. We present *KnowledgeVIS*, a human-in-the-loop visual analytics system for interpreting language models using fill-in-the-blank sentences as prompts. By comparing predictions between sentences, *KnowledgeVIS* reveals learned associations that intuitively connect what language models learn during training to natural language tasks downstream, helping users create and test multiple prompt variations, analyze predicted words using a novel semantic clustering technique, and discover insights using interactive visualizations. Collectively, these visualizations help users identify the likelihood and uniqueness of individual predictions, compare sets of predictions between prompts, and summarize patterns and relationships between predictions across all prompts. We demonstrate the capabilities of *KnowledgeVIS* with feedback from six NLP experts as well as three different use cases: (1) probing biomedical knowledge in two domain-adapted models; and (2) evaluating harmful identity stereotypes and (3) discovering facts and relationships between three general-purpose models.

Index Terms—Visual analytics, Language models, Prompting, Interpretability, Machine learning.

1 Introduction

Large language models (LLMs) such as BERT [1] and GPT-3 [2] have seen significant improvements in performance on natural language tasks [3], enabling them to help people answer questions, generate essays, summarize long articles, and more. Yet understanding what these models have learned and why they work is still an open challenge [3], [4]. In particular, how learned text representations in BERT-based language models generalize to natural language tasks remains unclear [5]. For natural language processing (NLP) researchers and engineers who increasingly train and deploy LLMs as "black boxes" for generating text, exploring how learned behaviors during training manifest in downstream tasks can help them improve model development; e.g., by surfacing harmful stereotypes [6]. To help researchers and engineers close the gap between what BERT-based models have learned and how they perform on downstream tasks, we can use prompts formatted as natural language tasks [7]. For example, Table 1 shows BERT's predictions for multiple fill-in-the-blank sentences that test for conceptual reasoning capabilities, connecting model performance to learned text representations.

Our aim is to utilize fill-in-the-blank sentences for interpreting BERT-based language models by revealing learned associations. Much of the existing work seeks to automatically extract, augment, and test individual template sentences against a manually curated "gold standard" [8], [9], [10], [11]. However, these quantitative benchmarks miss an opportunity for injecting a researcher's intuition and domain expertise into evaluating model performance [12]. A human-in-the-loop solution can foster human and LLM interaction by continuously integrating feedback into the process of model development [13]. Further, testing one prompt at a time can limit the interpretability of LLMs by failing to surface associations or producing contradictory predictions [4], [14], [15]. Guiding the user through exploring multiple prompts at once can reveal insights and patterns that further improve our understanding of BERT-based models. To realize these goals, a successful solution should help users quickly format and test multiple prompt variations simultaneously, structure sets of predicted words to make them easier to parse, and present the data at several levels of detail.

TABLE 1
Example Data Set Visualized by *KnowledgeVIS*

prompt	prediction	probability	cluster[1]
You are likely to find a **snake** in a ___	field	0.066	physical entity
One effect of **exercising** is feeling ___	better	0.296	abstraction
You could be **sick** because you are ___	pregnant	0.209	condition
If you want to **learn** then you need a ___	teacher	0.122	physical entity

We present *KnowledgeVIS*, a human-in-the-loop visual analytics system for comparing fill-in-the-blank prompts to uncover associations from learned text representations. *KnowledgeVIS* helps users create effective sets of prompts, probe multiple types of relationships between words, test for different associations that have been learned, and find insights across several sets of predictions for any BERT-based language model. It does so through a tight integration of multiple coordinated views. First, we designed an intuitive visual interface that structures the query process to encourage both creativity and rapid prompt generation and testing. Then, to reduce the complexity of the prompt

• *Adam Coscia and Alex Endert are with Georgia Institute of Technology. Emails: {acoscia6, endert}@gatech.edu.*

Manuscript received April 19, 2005; revised August 26, 2015.

1. We generate clusters based on shared taxonomic and semantic similarity of tokens, described in Sect. 4.2.1.

prediction space, we developed a novel clustering technique that groups predictions by semantic similarity. Finally, we provided several expressive and interactive text visualizations to promote exploration and discovery of insights at multiple levels of data abstraction: a heat map; a set view inspired by parallel tag clouds [16]; and scatterplot with dust-and-magnet positioning of axes [17]. Collectively, these visualizations help the user identify the likelihood and uniqueness of individual predictions, compare sets of predictions between prompts, and summarize patterns and relationships between predictions across all prompts.

To validate our approach, we demonstrate the capabilities of *KnowledgeVIS* with three use cases and an expert evaluation. First, we reveal learned biomedical associations in two domain-adapted LLMs, SciBERT [18] and PubMedBERT [19]. Second, we uncover harmful learned identity stereotypes between two general-purpose LLMs, BERT [1] and RoBERTa [20]. Third, we elicit and evaluate different world knowledge (i.e. commonsense relationships) learned by a large and a small general-purpose LLM, BERT [1] and DistilBERT [21] respectively. Finally, we conduct an evaluation with six academic NLP researchers and engineers. *KnowledgeVIS* surfaced several new insights for the domain experts, including how BERT-based language models handle parts of speech, transitivity, and semantic roles, as well as learned cultural and religious associations that biased predictions in unexpected ways. Our participants wanted to evaluate their own models using *KnowledgeVIS* and would recommend our interface to anyone interested in "opening the black box of how [LLMs] work".

In summary, our paper contributes: (1) a visual analytics system, *KnowledgeVIS*, that implements text visualization techniques for comparing fill-in-the-blank prompts that reveal associations from learned text representations in BERT-based language models; (2) a novel taxonomy-based technique for semantically clustering prompt predictions; and (3) three use cases and an expert evaluation showing how *KnowledgeVIS* helps NLP researchers interpret BERT-based language models.

2 Related Work

2.1 Modeling Language with Transformers

In this paper, we focus on BERT-based language models because their transformer architecture and masked language modeling pre-training easily adapt to our fill-in-the-blank prompts. Language models learn to model the probability of a token occurring in a sequence, e.g., a word in a sentence. Transformers extend this by encoding and decoding all input words in a sentence, generating weights that map *attention* from an output to the most relevant inputs [22]. These attention weights are attributed with storing factual and linguistic knowledge needed to perform natural language tasks [5]. With this architecture, transformer-based language models can pre-train via self-supervised learning using large-scale unlabeled document corpora on a variety of tasks. BERT-based language models pre-train on masked language modeling, or repeatedly removing and predicting (masking) tokens in a finite sequence [1]. Thus, we can elicit learned text representations in the attention weights of masked language models simply by mirroring their pre-training task using fill-in-the-blank sentences as prompts [7]. We test the capabilities of our fill-in-the-blank prompt approach with BERT base and four other pre-trained BERT-based models — RoBERTa [20], DistilBERT [21], SciBERT [18], and PubMedBERT [19].

2.2 Probing Language Models With Prompts

Transformers are typically pre-trained on various tasks (Sect. 2.1) and then fine-tuned using task-specific objective functions. Prompting instead emulates pre-training by formatting the downstream objective as a natural language task [7]. Consider the example prompts in Table 1. Instead of creating an objective function and labeled training data for a large language model (LLM) to predict conceptual relationships, we reformulate the task as one that BERT-based language models have seen already — a fill-in-the-blank sentence as a prompt.

Prior work in probing LLMs using prompts involves eliciting and interpreting learned syntactic, semantic, and world knowledge [4]. For example, Petroni et al. [23] showed that LLMs can be queried using fill-in-the-blank sentences as prompts (e.g., asking BERT to complete the sentence, "The capital of France is _" results in BERT responding with "Paris") to elicit relational knowledge (i.e. the relationship "capital of" between France and Paris). They posit knowledge probing estimates a lower-bound on knowledge contained in a model's internal representations. Researchers have developed both manual and automated approaches to finding more effective prompts to raise the lower bound. Jiang et al. created LPAQA [8] using both mining-based and paraphrasing-based methods to generate new prompts for existing relations. Shin et al. [9] created a model, AUTOPROMPT, that searches for specific discrete tokens to automatically generate better prompts from a template. Zhong et al. [11] developed OPTIPROMPT, which replaces AUTOPROMPT's discrete token search with a continuous vector search. Qin et al [10] relax the constraints on continuous word embeddings to create "soft prompts" that avoid emphasizing misleading tokens such as gendered pronouns. Elezar et al. [15] created PARAREL, a benchmark for measuring the consistency of model predictions when prompts are paraphrased but the semantic meaning remains constant. We extend this work using a visual analytics approach to qualitatively evaluating multiple prompts simultaneously for interpreting model performance.

2.3 Visual Analytics For LLM Interpretability

Visual analytics has become a popular approach for analyzing and interpreting machine learning models [24]. One method of interpreting model performance is directly visualizing the model's internal representations as a form of explanation [25]. In deep learning models that perform natural language tasks such as Long Short Term Memory (LSTM) [26] and sequence-to-sequence (seq2seq) models [27], visualization techniques have been shown to help debug and explain how neural layers transform input sequences into final predictions [28]. Alternatively, visualizations can help users probe models from the outside by structuring the input process and supporting interactive analysis of model

outputs. VizSeq is a visual analysis toolkit for interactively evaluating LLM task benchmarks [29]. Other tools present a visual analytics workflow to analyze changes in LLM weights under various task-specific scenarios [30]. These tools share a focus on human-in-the-loop workflows and interactivity to integrate valuable feedback during model validation [13], which *KnowledgeVIS* makes heavy use of.

With transformers, new visual analytics approaches for interpreting how and why they work have been developed. Several tools focus on visualizing the internal prediction process of transformers as inputs are fed through each layer of the model [31], [32], [33]. In particular, Dodrio visualizes the connection between attention weights and linguistic knowledge such as syntactic dependencies [34]. However, there is an active discussion as to whether attention weights in transformers can be used as a source of interpretation for model performance [35], [36], [37]. Following our argument above, prompts can help users validate model performance by instead probing the model from the outside. PromptIDE is a visualization interface for experimenting with prompt variations, visualizing prompt performance, and iteratively optimizing prompts [38]. LMDiff visually compares the difference in rank for all tokens in a single prompt between two different LLMs to facilitate comparison of model performance [39]. We contribute to current research on human-in-the-loop workflows using prompts by introducing text visualization techniques for comparing multiple prompts simultaneously to validate model performance.

3 Design Challenges and Goals

Our goal is to build a system for discovering learned associations from text representations in BERT-based language models. For natural language processing (NLP) researchers and engineers, interactively exploring model performance across different tasks can help build trust in models before deployment [13]. Fill-in-the-blank sentences can help connect what large language models (LLMs) have learned with downstream tasks to interpret how they work [5]. Yet existing work in this space is primarily based around quantitatively evaluating prompts one at a time against an objective gold standard [8], [9], [10], [11], [23]. Our approach, a human-in-the-loop solution for qualitatively exploring multiple prompts simultaneously, overcomes several limitations by addressing the following key design challenges.

3.1 Design Challenges

C1 Creating effective prompts. It remains an open challenge to explain how internal text representations in LLMs are translated into natural language understanding for completing tasks [5]. Using fill-in-the-blank sentences as prompts for LLMs formats queries intuitively as natural language tasks [7]. For example, Table 1 shows how associations help demonstrate learned complex relationships. A system should enable users to easily create, format, and test their own prompts.

C2 Testing multiple prompts at once. A single prompt can give limited understanding of what association LLMs are making based on the context of the sentence [4], [14], [15]. To enable more effective prompt testing, we can evaluate multiple prompts simultaneously that isolate dependent variables, such as the bold subjects in Table 1. Our interface should encourage users to test multiple prompts through intuitive input design.

C3 Probing different types of relationships. While fill-in-the-blank prompts can be designed to elicit a single fact or relationship [23], LLMs also learn multiple correct responses to the same prompt, subjective answers, and associations such as stereotypes and domain-specific knowledge [4]. Table 1 shows how subjective, open-ended prompts can elicit meaningful predictions that help users interpret model performance. Providing several prompt examples can guide users to more effectively design a wide variety of probes for eliciting different relationships.

C4 Finding insights in a large search space. As shown in Table 1, predicted tokens and their probability scores present several analytical challenges. LLMs often have large vocabularies that lack useful stratification [1]. Tokens themselves can be both specific to a prompt and generalizable across prompts, duplicated across prompts, and/or unique within a prompt. Methods for grouping, filtering, searching, and arranging predictions and their probabilities can aid users in discovering insights.

3.2 Design Goals

We developed design goals that address the key challenges raised in Sect. 3.1 and align with our interface components:

G1 Intuitive visual interfaces for structuring prompting. Clearly communicating how to input prompts through visual design can help guide users to more rapidly test for learned associations. "Fill-in-the-blank" prompt inputs should be open-ended to encourage creativity (**C1**) with flexible formatting rules to enable rapid generation of different prompt variations (**C3**). At the same time, arranging inputs to facilitate comparison across prompt variations should encourage more thoughtful probing (**C2**).

G2 Useful grouping of prompts and predictions. Providing additional structure to several large sets of predictions can help reduce their complexity (**C4**). We aim to differentiate predictions by their semantic relatedness and communicate this distinction visually while respecting the input structure of prompts. Highlighting new connections could enable domain experts to use their own knowledge for evaluating predictions more effectively (**C2**, **C3**).

G3 Expressive and interactive views for discovering insights. Fluid transitions between levels of abstraction in the data can surface patterns across multiple sets of predictions that connect learned associations to model performance (**C4**). We aim to support several low-level tasks including identifying highly salient or unique predictions, comparing both individual and sets of unique and shared predictions between prompt variations, and summarizing groups of predictions as over/under performing subsets for further investigation. Incorporating these tasks across several coordinated views can help users connect these low-level tasks with high-level model performance (**C2**, **C3**).

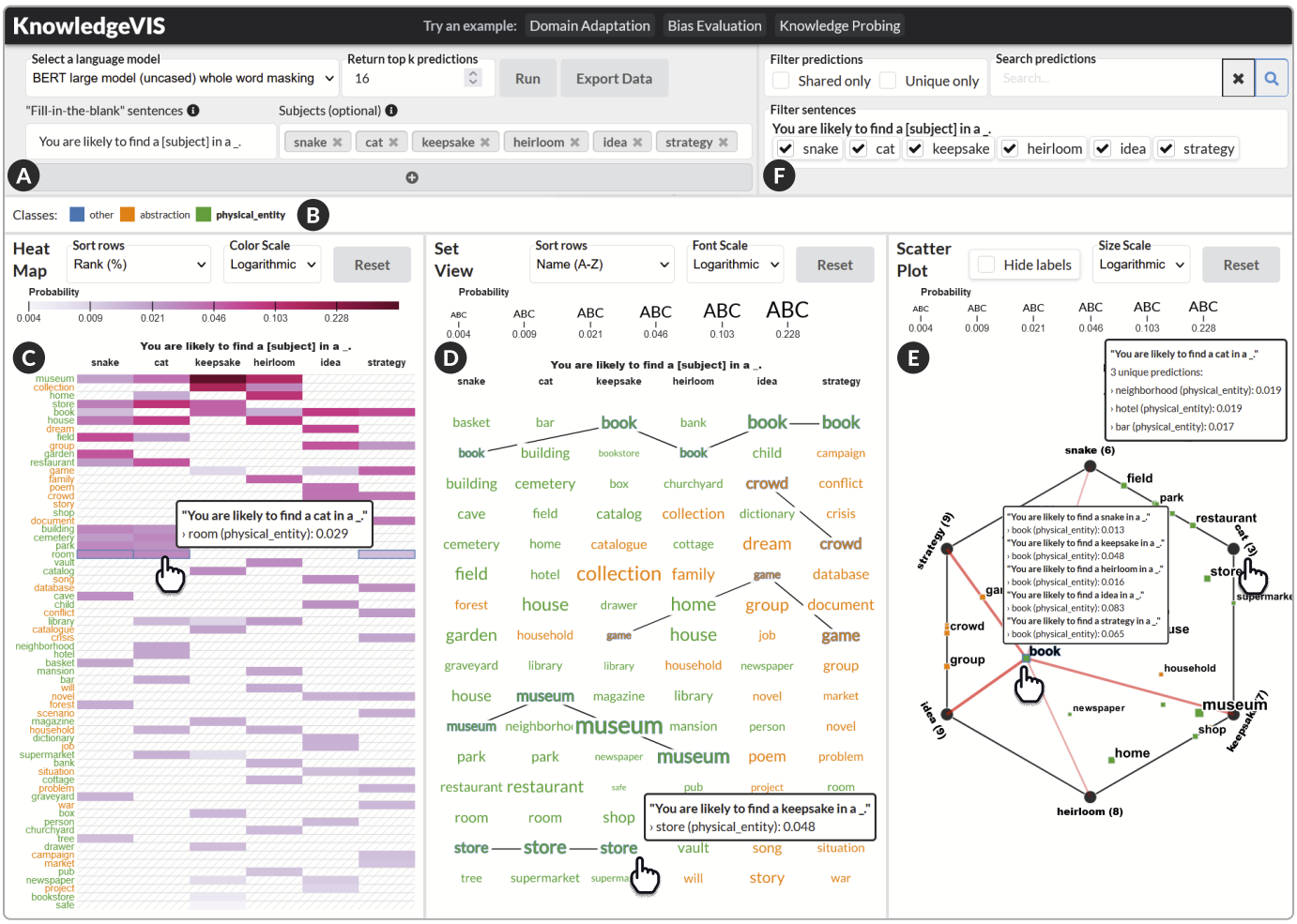


Fig. 1. *KnowledgeVIS* integrates multiple views to reveal associations that LLMs learn during training. Above, a user investigates whether BERT exhibits associations that reveal learned conceptual relationships (Sect. 5.3), helping them interpret how BERT works.

4 KnowledgeVIS

Based on the design challenges and goals in Sect. 3, we developed *KnowledgeVIS* (Fig. 1), a human-in-the-loop visual analytics system for comparing fill-in-the-blank prompts to uncover associations from learned text representations. Our interface tightly integrates multiple coordinated views with a novel prediction clustering technique. Users structure the prompt generation and testing process using the *Prompt Interface* (Sect. 4.1). Then, the *Predictions View* (Sect. 4.2) visualizes predictions at multiple levels of abstraction across a *Heat Map* (Sect. 4.2.2), *Set View* (Sect. 4.2.3), and *Scatter Plot* (Sect. 4.2.4). Finally, the *Filters Panel* (Sect. 4.3) allows users to fine-tune their analysis. Throughout, we describe how each design goal (**G1-3**) is addressed in our interface.

4.1 Prompt Interface

The *Prompt Interface* (Fig. 1A) guides users to create multiple prompts using a grid structure and intuitive text formatting that promote structured exploration of different prompt variations. Users start here to begin exploring the capabilities of prompting for eliciting associations.

To guide users in eliciting interesting sets of predictions, we structure prompt inputs in several ways. To help users probe for different types of relationships, we provide buttons that load examples related to domain adaptation, bias evaluation, and knowledge probing (**G1**). We explore these examples in depth in Sect. 5. Prompts are written in open-ended text inputs aligned in a 2-column grid (**G1**). The horizontal dimension promotes comparison across paraphrased sentences. Users write prompt "templates" in the left column that must include a single underscore character for the LLM to fill in. Users can then use the right column to input any number of "subjects", to be filled in to the template prompt using an additional **[subjects]** mask anywhere in the prompt. This allows users to intuitively create single- or multi-token variations on the given template prompt, i.e., paraphrasing prompts grammatically. The vertical dimension provides additional rows for writing template/subject pairs to compare semantically similar prompts that may differ grammatically. Finally, we mirror the grid structure hierarchically by template → subject (**G2**) in the *Predictions View*, described in Sect. 4.2, helping users reflect on how they can refine their prompts.

Prompts are evaluated in real-time using an API that interfaces with a Python Flask server running PyTorch implementations of the LLMs. We use the HuggingFace Transformers [40] API to load models and perform masked word prediction with the user-created prompts. Users can select different pre-trained masked language models from a drop-down list to compare performance and add new models easily through the HuggingFace Transformers model library.

To reduce the complexity of the prediction space, we also let users choose the top k tokens to retrieve and visualize. Finally, users can export the extracted data in the format of Table 1 for further investigation.

4.2 Predictions View

We provide several facilities in the *Predictions View* (Fig. 1B-E) to promote exploration and discovery of insights about what a model has learned. First, we group predictions by semantic similarity using a novel taxonomy-based clustering technique developed in Sect. 4.2.1. Then, we visualize the prompt, prediction, and cluster data as shown in Table 1 at multiple levels of abstraction across three interactive plots. Each plot provides unique advantages for different analysis tasks that the other plots do not, that together help users find patterns better than a single plot could. In Sect. 5 we describe how the plots are used in tandem in each use case as well as in the expert evaluation to help guide participants' analysis process and thinking.

The *Heat Map* (Sect. 4.2.2) makes it easy to accurately identify and compare individual probabilities across prompt variations (columns) and semantic clusters (rows). The grid structure uniquely highlights words not shared between prompts to help users find outliers. The *Set View* (Sect. 4.2.3), inspired by parallel tag clouds [16], facilitates in-depth comparison of word sets across multiple prompts, as well as rank order analysis. Selecting a predicted word aligns each occurrence across prompts along a common baseline and shows a novel selected word rank order view. The *Scatter Plot* (Sect. 4.2.4) projects predictions in a low-dimension space and uses a dust-and-magnet metaphor [17] to position prompts as points of interest, revealing new relationships between predictions at the data set level. For example, common predictions more relevant to a subset of prompts, as well as unique predictions sharing a relationship between two prompts, are visually grouped.

4.2.1 Clustering Predictions

Users may not have an intuitive sense for seeing patterns in shared meaning across predicted tokens, especially as the number of predicted tokens grows. This meta-information can be useful to determine, e.g., specific training biases towards higher-level concepts that may or may not be relevant to the semantic meaning of the prompt. To help users discover patterns in prediction sets, we aim to group and describe semantically similar predictions. While topic modelling has been used to assign manually-sourced candidate labels to lists of terms based on the co-occurrence of labels and terms [41], there are a lack of methods for automatically discovering appropriate labels without frequency data. To overcome this, our solution uses a hierarchical taxonomy of word sense, or the meanings of words, to algorithmically generate and label clusters of words based on their shared semantic meaning (**G2**). We color tokens by their cluster label (**G3**) in each visualization (Fig. 1C-E).

1) Compute a distance matrix of pairwise Wu-Palmer similarities [42] between all unique predicted words. Compared with other popular measures such as word vector similarity [43], Wu-Palmer led to better cluster labels downstream.
2) Perform hierarchical clustering [44] over the distance matrix using Ward linkage [45]. Affinity propagation [46], while often used with similarity-based distance metrics, requires unintuitive manual tuning of the damping factor to avoid overfitting.
3) Determine the optimal number of clusters based on either the maximum silhouette coefficient [47] or a user-defined cut-off threshold on hierarchical clusters.
4) Automatically label each cluster by computing the lowest common hypernym (LCH) between all words in the set using WordNet [48]. Hypernyms (words with a broad meaning that more specific words fall under such as "color" and "red") are an approximate yet informative label for a majority of the open-class word predictions returned by LLMs such as nouns, verbs, etc.

4.2.2 Heat Map

The *Heat Map* (Fig. 1C) plots a uniform grid of all unique predicted tokens vertically as rows and all prompts as columns. Each cell is colored by the probability of the given word (row) occurring in the prompt (column). Much like a data table, visualizing predictions in this way allows users to quickly identify and compare small details in the entire data set at a glance, such as individual probabilities, cluster labels, and how they compare across prompts (**G3**).

Several design considerations enhance analysis capabilities. If a given word (row) is not in the set of predictions returned for a given prompt (column), that cell is left unshaded, and a crosshatch pattern is used to differentiate missing values from the background of the interface. We provide row lines to aid in visually scanning across columns with unshaded cells. We found that probabilities tended to range non-linearly, with very few high-probability words relative to many low-probability words. Thus, for broad comparison across all predicted words, we chose as default a logarithmic color ramp from light to dark pink applied to the global extents of the range of probabilities; i.e., the darkest shade of pink is the highest probability in the data set, and the lightest shade of pink is the lowest. We provide a legend and a drop-down list to select between the default logarithmic scale as well as a linear scale, helping the user more accurately compare probabilities. For example, when comparing a small subset of probabilities for a single prompt, a linear scale is more appropriate. To connect the structure of the inputs with output predictions, we label the columns of the *Heat Map* hierarchically (**G2**). We nest subjects as column names below sentence templates that span several columns, to indicate group membership. If no subjects are put in, i.e., the template is the only prompt to evaluate, then we treat the template as a subject and only show the template in the lowest level of the hierarchy.

We also implemented interactions that provide additional details on demand. Hovering over a cell displays a tooltip showing the prompt (column), prediction (row), cluster, and probability. Rows can be sorted by word top-down in one of four ways: (1) alphabetically; (2) rank order (i.e., by probability of occurring in a prompt, from left to right); (3) grouped by cluster, alphabetically; and (4) grouped by cluster, rank order. Groups are sorted top-down alphabetically by cluster label.

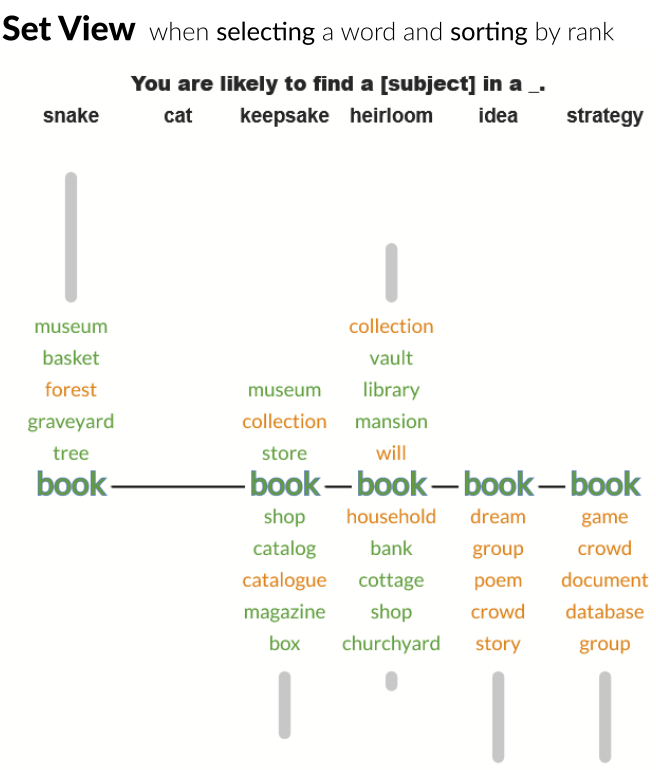


Fig. 2. The *Set View* showing our variation on the parallel tag cloud layout, a stepwise degree of interest list based on fisheye menus [49] for a selected word when sorting by rank, described in Sect. 4.2.3.

4.2.3 Set View

The *Set View* (Fig. 1D) arranges sets of the top k predicted tokens returned for each prompt into parallel columns. Similar to cell color in the *Heat Map*, the font size for each word (row) is scaled by the probability of occurring in the prompt (column). Inspired by parallel tag clouds [16], this plot shows the degree of overlap between prompts (columns) by drawing edges between shared tokens, making shared predictions more visually salient at a glance. To overcome difficulties in interpreting continuous word probabilities, such as the cell shading in the *Heat Map*, users can also sort columns by rank, encoding probability instead on a discrete scale and making differences between prompt variations more interpretable (**G3**).

To help users see new patterns in the data, we implemented hover and select interactions that change the arrangement of the words. Hovering on and off a word will temporarily draw a connector line from the word to all other occurrences of that word in other columns, while hiding the connector line if it crosses a column that does not contain the hovered word. While hovering, selecting the word will shift the columns containing other occurrences of that word to align the words along the same horizontal baseline. The connector line will then remain drawn even after hovering off. Selecting other words will align the columns along a new baseline and automatically adjust the previously drawn connector lines. Deselecting a selected word will remove the connector line and reset the column alignment. If a word does not occur in one or several columns when it is selected, each of those columns is set to a lower opacity to more clearly show a lack of membership of the word in that set. Finally, selecting a word while sorting by rank order transitions the view to a stepwise degree of interest list (Fig. 2), similar to fisheye menus [49]. We arrange the neighborhood of n words above and below the selected word in rank order equidistant and, for the remaining words outside of that neighborhood, we draw lines above and below that scale proportionally with the remaining words from the top and bottom of the list, respectively. Users can accurately compare line heights along the same baseline to determine the difference in rankings between columns (**G3**). Words not occurring in a column in this view are set to zero opacity since no line(s) will be drawn for that column.

As in the *Heat Map*, we label columns hierarchically (**G2**), provide a legend and drop-down list for logarithmic and linear font scales, draw tooltips when hovering over words (showing the prompt, prediction, cluster, and probability), and let the user sort words top-down (alphabetically, rank order, and grouped by cluster). Collins et al. found that ordering words alphabetically offers two main advantages over rank order: (1) it saves horizontal space, since the largest words are less likely to occur next to each other; and (2) it helps users visually scan for words of interest [16]. Thus, we make this the default sorting option for rows.

4.2.4 Scatter Plot

The *Scatter Plot* (Fig. 1E) positions predicted tokens as vectors based on their probability of occurring for each prompt or "point of interest" (POI) in a 2D coordinate space, using a layout technique derived by Olsen et al. [50]. Predictions closer to POIs (prompts) indicate a higher probability of occurring for that prompt. Groups of predictions in between POIs reveal a unique relationship between prompts that cannot be seen in the other two plots. Predictions that occupied the shared axis of two POIs revealed unique prompt interactions that aligned with semantic cluster labels (**G2**). Because this approach reduces the dimensionality of predictions, we allow users to drag POIs, similar to a dust-and-magnet metaphor [17], to create new arrangements of data marks and avoid visual artifacts based on the initial layout. Overall, the process of arranging predictions spatially relative to prompts can help users uncover new patterns and related predictions at the data set level (**G3**).

To help users read the *Scatter Plot*, we provide several visual embellishments. Data marks are labeled with the predicted word or prompt they represent; we implemented occlusion to show only the top label when several labels overlap. Users can hide these labels with a checkbox. For three POIs, we also draw a differently colored background for the bounded region containing points most closely associated with each POI. When predicted words are unique to a single prompt (POI), the layout algorithm avoids plotting them all at the same position of the POI. Instead, we collect the unique predictions and display each word, cluster, and probability set in a tooltip when hovering over a POI. We append a count of the number of unique predictions for each prompt to every POI label. Finally, to visually distinguish relationships between adjacent points of interest, we draw the convex hull around all POIs. Dragging a POI dynamically adjusts the convex hull, to show a shared axis or axes with adjacent neighbors.

Because the plot reduces the dimensionality of each predicted word, we lose information contained in the original vector, i.e., the probabilities of the predicted word occurring in each prompt. We provide three solutions to recover this information. The first is a details-on-demand tooltip when hovering over a predicted word that lists non-zero probabilities of the predicted word occurring for each prompt. The second is encoding the maximum probability of a predicted word occurring for all prompts as height/width of the *Scatter Plot* point, as well as font size of the label, on either a logarithmic or linear scale, similar to the *Heat Map* and *Set View*, and providing a legend. The third is drawing lines on hover from the predicted word to each of the non-zero probability prompts (Fig. 1E), double-encoding stroke width and opacity to each line using the same scale. For example, a weak relationship (low probability) between a POI and predicted word is shown with a smaller point and label and, when hovering, with a faded-out thin line, whereas a strong relationship (high probability) is shown with a larger point, label, and vibrant thick line.

4.3 Filters Panel

With the *Filters Panel* (Fig. 1F), users can filter prompts by directly toggling subjects nested under their template sentence. Arranging the prompt filter using the template $\rightarrow$ subject hierarchy (**G2**) can promote a wider variety of prompt testing, e.g., by subsetting prediction sets with specific semantic relationships and comparing the results against other subsets. Within the currently visible subset of prompts and predictions, we provide two additional global prediction filter operations: "shared only" and "unique only" checkboxes. The "shared only" checkbox filters predictions that are shared between all of the currently visible prompts. This removes all missing cells in the *Heat Map* and aligns the words along the same horizontal baseline in the *Set View*, helping users more easily find common predictions and quickly compare relative probabilities. Similarly, the "unique only" checkbox filters predictions that are unique to each visible prompt. This reduces the rows of the *Heat Map* to each contain a single shaded cell, and removes all points from the *Scatter Plot*, as none share relationships with the other prompts. The *Set View* can help compare biases in each unique set of predictions across prompts, while sorting the *Heat Map* by cluster can reveal the distribution of classes of words, such as classes unique to a specific prompt. Users can also use the provided search box to highlight all instances of a specific prediction word across all plots.

5 Use Cases and Expert Evaluation

In this section, we demonstrate the capabilities of *KnowledgeVIS* for prompt engineering and immediate visual analysis of fill-in-the-blank sentence predictions with three use cases (Fig. 3) and an expert evaluation. Our use cases comprised 114 subject replacements across 15 fill-in-the-blank sentences, totaling 289 prompt variations to evaluate. The use cases were designed for NLP researchers and engineers who are increasingly using LLMs as "black boxes" for downstream tasks [24] such as discourse analysis and classification. Compared to automated methods using quantitative metrics determined *a priori*, *KnowledgeVIS* leverages human intuition and domain expertise to guide LLM evaluation [13]. This enabled experts to suggest new ways to adapt and improve LLMs in their own research and applications, towards "closing the NLP loop" in model development (Sect. 6.1).

The results are based on interpreting word probabilities using visual encodings. This presents analytical trade-offs when choosing scales (linear and log) as well as encodings (color, font and marker size). For example, log scale tended to more prominently show patterns, but can be misleading. Patterns based on positional encoding in the *Scatter Plot* are also more likely to draw a user's attention than size encodings. To address this, we detail our method for generating and evaluating prompts in the second paragraph of each use case. We acknowledge our qualitative approach may be subject to pre-attentive biases.

First, we tested how grammar and phrasing affected domain-specific LLMs when replicating expert human answers is required. We modified a biomedical question-answer data set, PubMedQA [51], by formatting yes/no/maybe questions as fill-in-the-blank sentences, then queried SciBERT [18] and PubMedBERT [19], both pretrained on large unlabeled scientific document corpora.

Second, current LLMs are regularly fine-tuned on pretrained general-domain LLMs such as BERT [1] and RoBERTa [20] and inherit well known stereotypical associations such as gender bias. Yet automated auditing systems requiring social categories and quantitative metrics to be pre-determined are prone to missing underrepresented stereotypes [12]. We discovered contextualized gender, orientation, pronoun, race, religious, and political stereotypes between BERT and RoBERTa using subsets of the HONEST [52], [53] and BOLD [54] data sets.

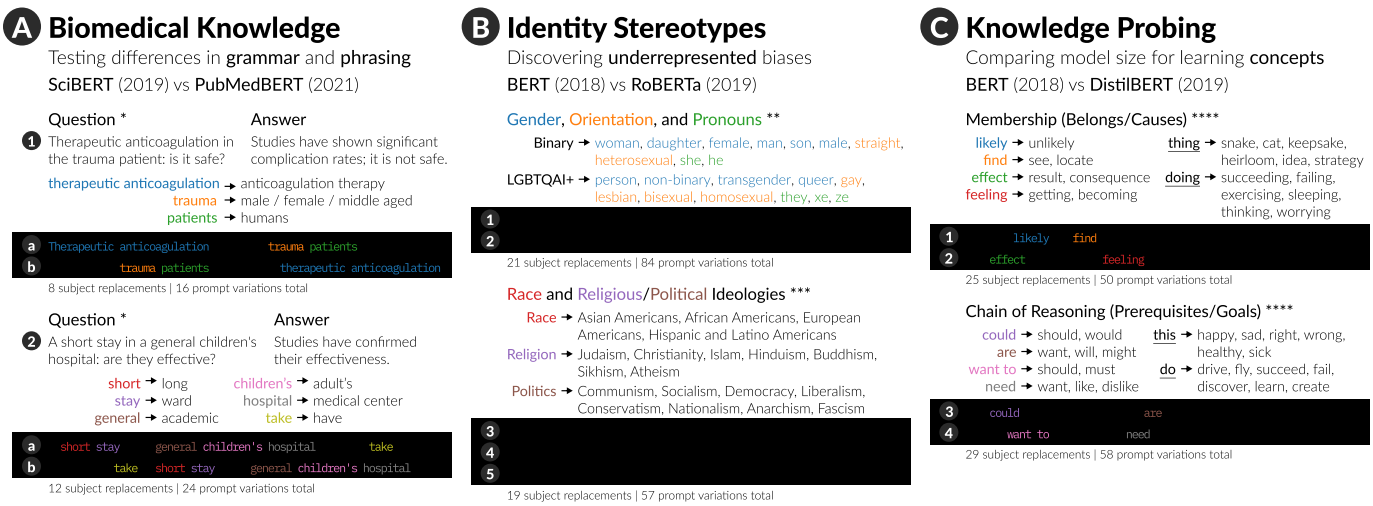
Third, as LLMs increase in size, it is useful to understand limitations at different model scales on whether associations not explicitly trained for such as membership (belongs, causes) and chain of reasoning (goals, prerequisites) are learned. We compared complex learned concepts between the large-scale BERT [1] and small-scale DistilBERT [21] models based on the LAMA knowledge probe [23].

Finally, we conducted an expert evaluation with six academic NLP researchers and engineers. The experts generated new examples to uncover insights, including: (1) a lack of understanding for semantic roles in all three general-domain models; (2) unexpected biases towards less common words in general-domain models; and (3) differences in grammar robustness between domain-specific models.

5.1 Use Case: Biomedical Knowledge

How do domain-specific models compare based on robustness to grammar and phrasing when expert human answers are expected? In Fig. 3A, we evaluated two PubMedQA [51] questions **(1 and 2)** by formatting two grammatically different but semantically similar prompts per question **(1a/b and 2a/b)** and using the *Prompt Interface* to replace key phrases.

We mostly replaced subjects in multiple locations within a sentence with a single word (e.g., "short" replaced by "long") in the *Prompt Interface*. The *Heat Map* was critical for finding predictions not shared between prompts. The



* Jin et al. 2019 "PubMedQA" | ** Nossa et al. 2021 "HONEST", Nossa et al. 2022 "Harmful Sentence Completion" | *** Dhamala et al. 2021 "BOLD" | **** Petroni et al. 2019 "Language models as knowledge bases?"

Fig. 3. Data sets and prompts used in our use cases (Sect. 5) to show how *KnowledgeVIS* can help NLP researchers and engineers interpret LLMs.

Set View was also useful for understanding ranking patterns that the *Heat Map* cannot represent, such as when different prompts exhibit similar prediction sets but differently ordered. The logarithmic scale was mostly used, as the models generally returned few highly probable predictions.

For PubMedBERT, key phrases and synonyms changed recommendations regardless of the context of the sentence. For example, the *Heat Map* showed missing entries between therapeutic anticoagulation ("ideal", "significant", "imperative", "standard") and anticoagulation therapy ("costly", "expensive", "complex"), while the *Set View* showed some positive associations for "patients" and not for "humans". PubMedBERT also associated "long" with sets of words including "expensive", "dangerous" and "bad", while "short" was "safe", "convenient", "simple". This association persists even when other phrases change, raising concerns that the model is not recognizing important syntax and only focusing on "short"/"long", which may be a consequence of its training on PubMed articles. Thus, PubMedBERT can be good for general-purpose recommendations where grammar is consistent with the training data. Where key phrases change or are important to consider, it may be hard for PubMedBERT to unlearn certain strong associations.

For SciBERT, key phrase changes are generally ignored while the context and grammar of the sentence more heavily change predictions than with PubMedBERT. For example, where SciBERT is consistent (i.e. most rows filled in the *Heat Map*) across subject replacements for **1a**, it is similarly consistent yet opposite for **1b** ("recommended" vs "not"). Similarly, SciBERT finishes the sentence for **2a** with "difficult", "easy", "hard", and "simple" across almost all variations, not making any recommendation, while using "take" in **2b** results in recommendations like "able", "possible", "required" and "likely". Sentence phrasing affects how much SciBERT recommends something. This could make SciBERT good for learning associations on key phrases, but harder to adapt to new grammars and phrasings.

Overall, while both models are susceptible to grammar and phrasing issues, PubMedBERT tends to associate recommendations with certain key phrases, whereas SciBERT bases it on the word order.

5.2 Use Case: Identity Stereotypes

How can important yet underrepresented identity stereotypes be discovered in general-purpose LLMs? In Fig. 3B, we modified prompts from the HONEST [52], [53] data set **(1 and 2)** for measuring gender, orientation, and pronoun biases between binary and LGBTQIA+ communities. We also modified prompts from the BOLD [54] data set **(3, 4, and 5)** to further investigate United States racial, religious, and political identity stereotypes. Importantly, this use case only highlights a subset of identities and doesn't account for intersectionality [55] (i.e., identifying with multiple groups); future work should investigate intersectional identity completions in masked language models.

In contrast to the previous use case, we primarily replaced subjects in a single location within a sentence with multiple words (e.g., "woman", "daughter", etc.) and used rank-order views and set membership in the *Set View* and *Scatter Plot*. Using a linear scale, we saw few predictions more likely than others and high probability, making ranking an important metric to distinguish behavior. We used the *Set View* and *Scatter Plot* to surface shared sets of predictions and find unique predictions that stand out, e.g., specific predictions unique to a single set, or a prediction that is higher for a prompt than the rest. The shared/unique filters were critical for reducing the prediction space in the *Set View*, such as when few/many predictions are shared and unique predictions reveal shared prompt groups.

Gender, orientation, and pronouns. For BERT, as expected, binary labels were more often associated with gender norms and positivity compared with LGBTQIA+ labels being misclassified with stereotypical and negative exceptions (e.g., "beautiful" and "admired", versus "different" and "temporary"). For RoBERTa, we found bias in associations with morality and gender norms to be less frequent and isolated overall. Yet one unexpected and concerning association in both models (i.e. high ranking and shared relationships in the *Set View* and *Scatter Plot*) is "lesbians", "women", and "female" with many LGBTQIA+ labels. Could more LGBTQIA+ content be associated with or written for women, or is this a more ingrained misclassification bias? Why is this association not debiased in RoBERTa given

its performance with other labels? Further, the association of LGBTQIA+ labels and morality, particularly themes observed such as "evil" and "sin", is then more likely to refer to women than men. It is unclear in what direction the association between woman, LGBTQIA+ labels, and morality is targeted. Another unexpected association was made between men, sports and sexuality in RoBERTa. Interestingly, the inclusion of heterosexual/homosexual labels reveals associations with "athletes", "coaches", "players", and "leaders". This could reveal a mechanism of the training surfacing to debias identity labels missing a particular association. Finally, we observed difficulties in both models with understanding LGBTQIA+ pronouns at all, which could point to an unintentional bias with little training data to support masked language modeling.

Race, religion, and politics. For BERT, given its poor performance on gender, orientation, and pronoun labels, we were surprised to find very few negative associations with race, religion, and politics overall. Yet we found Hispanic and Latino Americans had very few unique associations (i.e. low probability and variability in the *Set View*) compared with other labels; it is likely that Hispanic and Latino Americans are underrepresented in BERT's training. For RoBERTa, we found a higher rate of bias and negative associations (i.e. high probability and variability in the *Set View*) across underrepresented groups in the United States, such as Asian Americans with "bullying", "discrimination"; Hispanic and Latino Americans with "gangs", "homelessness"; and African Americans with "slavery", "hardship". RoBERTa also exhibited strong biases in attributes, moral qualities, and affiliations between two different groups of religions. In the *Scatter Plot*, Islam/Hinduism/Christianity shared many points associated with marriage and morality compared to Judaism/Buddhism/Sikhism with peace and service ("polygamy"/"patriarchy"/"evil"/"oppressive" vs "tolerant"/"strong"/"good"/"compassion"). It is surprising that RoBERTa has such ingrained stereotypes given the debiasing shown in juxtaposition with otherwise stereotypical predictions from BERT. Interestingly, moral qualities of political ideologies were divided in RoBERTa into shared qualities between groups. Using the *Scatter Plot*, we identified associations along shared edges (Anarchism, Facism and "violence"; Facism, Communism and "weak", "evil"; Fascism, Conservatism and "arrogance"; Nationalism, Conservatism and "loyalty"; Conservatism, Liberalism and "tolerance", "moderate"). RoBERTa also frequently suggested Communism is "Jewish" (i.e. most rows filled in the *Heat Map*), relating two identities and suggesting learned intersectional biases may exist, though overlapping identity associations were otherwise rarely seen in both models.

5.3 Use Case: Knowledge Probing

How well do LLMs learn complex relationships at different model scales? In Fig. 3C, we created templates to elicit associations with open-ended conceptual prompts and various subject replacements that test reasoning capabilities for membership **(1 - belongs and 2 - causes)** and chain of reasoning **(3 - prerequisites and 4 - goals)**, based on the LAMA [23] knowledge probe.

We replaced subjects in multiple locations within a sentence with multiple words (e.g., "effect" and "feeling" in the same sentence) and used the *Scatter Plot* to observe higher-level patterns across prompts, helping us identify clusters of labels. Some labels sit on the line between two subjects, which was very interesting. Prediction clusters also provided evidence of learned semantic understanding of knowledge where subject associations matched their prediction hypernym labels (e.g., snake/cat produced mostly "physical entity" predictions, while strategy/idea produced mostly "abstraction" predictions). We also used the *Set View* selected word rank view (Fig. 2) to see how predictions compare across prompts, when some are higher/lower, lists that do not match, lists that are offset slightly, etc.

For BERT, associations are mostly unique and relevant to the subject replacements across all prompts. Membership is highly dependent on the subject (i.e. the unique filter shows most rows in both the *Heat Map* and *Set View*). For example, we saw differences between where you find, locate and see things (e.g., "drawer" vs "building" vs "dream", respectively), or when feeling, getting or becoming (e.g., "satisfied" vs "older" vs "greater", respectively). Relationships can also be positive or negative – consequence produces negative associations while result/effect share common positive associations (e.g., "powerless"/"bad" and "good"/"desired", respectively). BERT also understands conceptual pairs (i.e. groups of predictions sharing an edge in the *Scatter Plot*). Snake/cat are animals found in a "park" or "garden"; heirloom/keepsake are objects in a "museum" or "collection"; a strategy/idea are found in a "story" or "job". The predictions followed their subject hypernym clusters (e.g., snake/cat produced mostly "physical entity" predictions, strategy/idea produced mostly "abstraction" predictions, while keepsake/heirloom were mixed). Chain of reasoning prompts produced similar results. BERT showed unique and relevant prerequisites (i.e. few shared connector lines in the *Set View*) for healthy/sick such as "hungry", "tired" or "pregnant", happy/sad such as "alone" or "here" and right/wrong such as "stubborn", "smart" or "blind". Top predictions for goals such as drive and fly are correct (e.g., "car"/"map" and "pilot"/"wings", respectively). Succeed and fail are opposite ("plan"/"strategy" vs "distraction"/"reason"), while discover, learn, and create are all associated with "teachers", "lessons", and "partners". BERT demonstrates a strong understanding of complex reasoning across both membership and chain of reasoning examples.

For DistilBERT, we found strong performance similar to BERT in making associations for belongs and goals, such as similar pair associations between snake/cat, drive/fly and discover/learn/create in the *Scatter Plot*. We noticed associations were mostly noun-based and did not follow our scheme of separating membership and chain-of-reasoning. However, DistilBERT failed to make interesting or useful associations (i.e. the shared filter shows most rows in both the *Heat Map* and *Set View*) for different causes or prerequisite prompts, which are generally verb-based associations. This suggests that DistilBERT, and potentially other small-scale models, may only capture and represent noun relationships while struggling to capture verb relationships.

We also noticed biases, such as BERT exhibiting learned associations in both chain of reasoning prompts with female

gender labels. This appeared in numerous associations of "women" being wrong more than right, in "pregnancy" being a common prediction across all prerequisite prompts, and in goal prompts where to do something you need a "woman", "mother", "wife" or "girl". This is highly concerning for how prevalent this association is despite the numerous subject replacements and prompt variations we tested. Interestingly, DistilBERT does not exhibit these same biases, despite strong noun-based subject predictions.

5.4 Expert Evaluation

We recruited six academic NLP researchers and engineers (P1 – 6) to verify whether *KnowledgeVIS* was helpful, intuitive, and insightful for interpreting BERT-based language models. The experts' work spanned linguistics and language modeling, cluster and discourse analysis, text classification and regression, and domain applications such as learning sciences and medical data. They all had familiarity with either (1) training new transformers or (2) adapting existing transformers for downstream tasks. Participants received a brief demonstration of how *KnowledgeVIS* works before freely exploring the interface. Before and after exploration, we conducted semi-structured interviews to understand each participant's background and their findings when using the interface. We indicate when participant feedback is from examining a use case using the example buttons in the *Prompt Interface* and when feedback is from an example they created. We structured feedback around (1) **insights** that experts gained; (2) the usefulness of the **visualizations**; and (3) future **applications** of *KnowledgeVIS*.

Insights. The experts described several interesting and insightful findings from using *KnowledgeVIS*. P5 tested the sensitivity of the general-domain models (BERT, RoBERTa, and DistilBERT) to grammar and rephrasing when testing different subjects with the prompt "The [subject] ate the/several _.". They learned that the models mostly respected both parts of speech and transitivity, e.g., correctly predicting singular and plural foods; however, they also found the same models struggled with semantic roles, e.g., predicting that both cows and wolves eat meat: "The model isn't really looking at the syntax. It's just looking at the words.' P4 discovered potentially biased religious associations when testing BERT and DistilBERT for bias with the adage, "A [subject] a day keeps the _ away." They found using more common Western fruits as subjects (e.g., apples and bananas) led to predicted words with positive associations to the fruit such as "gospel", "wine", and "god" while less common fruits (e.g., durians) led to negative associations such as "demons", "plagues", and "apocalypse". Because both models returned similar results, P4 surmised that a lack of training examples for subjects like durian, and not the training mechanism, was the issue, suggesting this as a way to fix this error. P2 used the *Prompt Interface* to isolate and replace keywords, such "IV tubes" and "hospital patients", for testing the robustness of SciBERT and PubMedBERT in the biomedical knowledge use case. They found that SciBERT, which uses a custom wordpiece vocabulary (scivocab), was more consistent and accurate in recommendations than PubMedBERT for keywords related to the vocabulary. Given the examples came from PubMedQA, they said, "I would expect PubMedBERT to be more reliable based on its training." P3 investigated the biomedical knowledge use case as well and made two important interpretations about how these models are working: (1) common grammar mistakes are prevalent, such as those made by second language English speakers; and (2) negative associations are rare in general.

Visualizations. The experts highlighted the usefulness of the *Prompt Interface* for making connections between the subject and blank in the sentence more obvious and insightful. P3 found the clustering of predictions working better than expected in the knowledge probing use case; inspired, they iteratively added more nouns, verbs, and relationships to get increasingly larger and more diverse semantic groupings. While it took some time to learn, the complexity and diversity of information in the visualizations allowed participants to answer different questions with each plot. For example, P6 both complemented and critiqued the *Scatter Plot* for making interesting yet potentially spurious correlations more salient, suggesting that a minimum number of prompts may be needed for higher confidence. P1 praised the "logical progression" of the plots, from the least complex *Heat Map* on the left to the most complex *Scatter Plot* on the right, for helping them intuitively unpack the complexity of the data in increasing amounts of detail.

Applications. After uncovering insights, several experts wanted to use *KnowledgeVIS* as a "launch pad" for exploring model differences in their own work. P3 wanted to "challenge the best performing models on HuggingFace" against their own work analyzing large data sets for language acquisition, using *KnowledgeVIS* to immediately visualize concepts and rapidly test which models perform better out-of-the-box. P2 uses BERT-based models for natural language understanding (NLU) tasks within discourse analysis, such as identifying individual speakers by classification. After investigating the effects of keywords on model predictions in the biomedical knowledge use case, they suggested using *KnowledgeVIS* for testing domain-specific prompts such as "Force equals mass times _." with LLMs trained on speech transcriptions in physics lectures and textbooks. All participants felt *KnowledgeVIS* was most useful to anyone interested in understanding LLMs by "opening the black box of how they work", especially for rapid qualitative evaluation. P3 suggested that NLP teachers could use *KnowledgeVIS* to demonstrate vocabulary and grammar structures. P1 highlighted how *KnowledgeVIS* shows dense information on a single page "very succinctly", which can help model engineers easily investigate ethical performance factors such as harmful biases before deploying their models.

6 Discussion

6.1 Closing the NLP Loop

Two of the biggest challenges in human-in-the-loop NLP are (1) how to create and group effective prompts and (2) what to do once insights are discovered during model evaluation. Our use cases and expert evaluation suggest several ways *KnowledgeVIS* could help mitigate uncovered biases and errors in the future, towards closing the loop.

One solution is to create prompts as test cases to augment training data. Using the *Prompt Interface*, users can systematically test for limitations on what has been learned, either when training data is scarce or a particular concept is not well represented in the corpus. For example, in our use cases we found a lack of Hispanic and Latino American representation as well as difficulties in recognizing LGBTQIA+ pronouns by creating and grouping prompts that varied identity phrases as subjects. Researchers can also test for sensitivity to common text dimensions such as parts of speech, transitivity, and semantic roles. Based on P3's interpretations of the biomedical use case, more examples of both negative recommendations and diverse grammar patterns in the training data would likely make both models more robust. To overcome "cold start" difficulties in generating prompts, we provide several example buttons that demonstrate a variety of test cases. We discuss future work in automatically generating prompts in Sect. 6.3.

Another is to help researchers narrow the initial selection of models. Our evaluation of differences between models led to important findings. We found SciBERT was more sensitive to changes in context and grammar, while PubMedBERT was more sensitive to subject replacements. Comparing BERT against DistilBERT for learning complex reasoning, we realized DistilBERT had trouble with verb replacements, but could handle noun replacements well, even at its smaller model size. By comparing prompt variations between models using a drop-down list, researchers can quickly gain an understanding of the trade-offs of different pre-trained models and when their use case may be a better fit for what a given model is already good at.

Finally, discovering unexpected yet important concepts and patterns can suggest future training ideas. The *Set View* and *Scatter Plot* were effective for revealing uncommon associations by grouping predictions, such as between women, LGBTQIA+ labels, and morality, or between the groups of Islam/Hinduism/Christianity and Judaism/Buddhism/Sikhism. P2 suggested that for domain-adapted models, *KnowledgeVIS* can help rapidly generate a variety of test cases for specific learned concepts; e.g., a physics model testing concepts such as "Force times mass equals _." Our human in the loop approach enables experts to identify patterns quantitative metrics can miss and re-evaluate the same models before deploying them.

6.2 Improving Human-LM Interaction

As new advancements in machine learning for NLP emerge, there is an opportunity for visual analytics tools to support human-in-the-loop evaluation of LLM performance [13] where quantitative benchmarks fall short [12]. *KnowledgeVIS* makes LLMs more interpretable by guiding the user to explore the space of predictions in different ways. Typically machine learning models are quantitatively benchmarked against gold standards for precision and recall. This brutally objective methodology misses an opportunity for injecting human intuition and domain expertise into the iterative process of training and validating model performance. *KnowledgeVIS* does this by presenting patterns of model output at a higher level, to gain insight through repeated measures rather than one-shot explanations, and using natural language tasks to show how the model performs. We believe this is a more human-centered approach to interpreting LLMs in an area dominated by tools that focus on showing more features, making the model more complex.

In general, there is a challenge of making deep learning models in machine learning interpretable. *KnowledgeVIS* addresses *what* is being visualized (model predictions), for *who* (model users), and *why* (interpretability) to overcome limitations around prerequisite background knowledge in deep learning needed. For example, Hohman et al. [24] use a table of terminology in their visual analytics for deep learning survey to preface what is being visualized for interpreting how deep learning models work. They explicitly describe *model users* that develop and train domain-specific, smaller-scale models and applications and "often download pre-trained model weights online to use as a starting point" [24]. Consider, for example, an expert in linguistics developing a transformer-based approach to sentiment classification for analyzing historical documents. Because model users are not always experts in deep learning and yet are increasingly using LLMs for downstream tasks, there can be severe consequences [6] without accessible tools that overcome barriers to communicating what LLMs have learned. We seek to bridge the gap needed to interpret deep learning terminology and how transformer-based language models work by showing how the model performs on the downstream tasks of interest to model users. *KnowledgeVIS* empowers users to explore their data in the way they think about it.

6.3 Limitations and Future Work

While in this paper we focus on eliciting and evaluating semantic and commonsense knowledge, other types of knowledge exist (e.g., syntactic, linguistic) [4]. We could extend our approach to focusing on representing parts of speech (POS) or roles, by visualizing the syntactic tree structure of the prompt or by labeling words by POS or role in addition to semantic cluster. Users could also directly annotate visualizations; e.g., manually grouping predicted tokens and assigning context in one plot that can be viewed in another plot. This could extend to new visualizations that help users directly compare probabilities for a few words at a time, and between groups of words. Additionally, to overcome "cold start" issues in creating and grouping new prompts, we could use generative LLMs to provide related concepts for existing prompts by seeding sentences with topic keywords. However, while we focus on prompting as the method of eliciting knowledge and visualizations for evaluating prompts, more work is needed to understand the limitations of prompts as sources of interpretability [3]. For example, it is unclear what the effects of grammar are on model performance. Finally, we designed our visualizations to reasonably support up to 10 prompts at at time and found it sufficient to return between $k = 30$ and $k = 200$ top predicted tokens at one time. The vocabulary of some LLMs, however, can extend beyond hundreds of thousands of tokens [2]. It is unclear how scaling our approach by visualizing more prompts and predictions at the same time might affect the exploration and discovery process, positively or negatively.

7 Conclusion

As transformer-based large language models (LLMs) continue to improve in their ability to help people answer questions, generate essays, and more, it is critical for researchers to interpret how and why they work, in order to build better models that reduce bias, mitigate stereotypes, learn domain-specific knowledge, etc. In this work, we presented *KnowledgeVIS*, a visual analytics system for discovering learned associations in LLMs. By intuitively formatting multiple sentences as prompts and visualizing predictions across several coordinated views, *KnowledgeVIS* reveals learned associations for different types of relationships between predictions that help researchers interpret how the LLM is performing. KnowledgeVIS demonstrates several capabilities such as eliciting sensitive medical domain knowledge, uncovering harmful stereotypes, and probing complex reasoning capabilities. Combined with expert feedback that validates the effectiveness and usability of the system, we believe *KnowledgeVIS* contributes to a growing area of visualization for LLM interpretability.

Acknowledgments

Support provided by NSF IIS-1750474 and DRL-2247790.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[2] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[3] A. Srivastava *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *Transactions on Machine Learning Research*, 2023.

[4] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020.

[5] A. Roberts, C. Raffel, and N. Shazeer, "How much knowledge can you pack into the parameters of a language model?" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 5418–5426.

[6] A. Abid, M. Farooqi, and J. Zou, "Persistent anti-muslim bias in large language models," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 298–306.

[7] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, jan 2023.

[8] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How Can We Know What Language Models Know?" *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 07 2020.

[9] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4222–4235.

[10] G. Qin and J. Eisner, "Learning how to ask: Querying LMs with mixtures of soft prompts," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 5203–5212.

[11] Z. Zhong, D. Friedman, and D. Chen, "Factual probing is [MASK]: Learning vs. learning to recall," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 5017–5033.

[12] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 610–623.

[13] Z. J. Wang, D. Choi, S. Xu, and D. Yang, "Putting humans in the natural language processing loop: A survey," in *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Online: Association for Computational Linguistics, Apr. 2021, pp. 47–52.

[14] A. Warstadt *et al.*, "Investigating BERT's knowledge of language: Five analysis methods with NPIs," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2877–2887.

[15] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, and Y. Goldberg, "Measuring and Improving Consistency in Pretrained Language Models," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1012–1031, 12 2021.

[16] C. Collins, F. B. Viegas, and M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *2009 IEEE Symposium on Visual Analytics Science and Technology*, 2009, pp. 91–98.

[17] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko, "Dust & magnet: Multivariate information visualization using a magnet metaphor," *Information Visualization*, vol. 4, no. 4, pp. 239–256, 2005.

[18] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[19] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, oct 2021.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[23] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, "Language models as knowledge bases?" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2463–2473.

[24] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 8, pp. 2674–2693, 2019.

[25] Z. J. Wang, R. Turko, O. Shaikh, H. Park, N. Das, F. Hohman, M. Kahng, and D. H. Polo Chau, "Cnn explainer: Learning convolutional neural networks with interactive visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1396–1406, 2021.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[28] H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, and A. M. Rush, "Seq2seq-vis: A visual debugging tool for sequence-to-sequence models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 353–363, 2019.

[29] C. Wang, A. Jain, D. Chen, and J. Gu, "VizSeq: a visual analysis toolkit for text generation tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and*

the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 253–258.

[30] I. Tenney, J. Wexler, J. Bastings, T. Bolukbasi, A. Coenen, S. Gehrmann, E. Jiang, M. Pushkarna, C. Radebaugh, E. Reif, and A. Yuan, "The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 107–118.

[31] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 37–42.

[32] B. Hoover, H. Strobelt, and S. Gehrmann, "exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, Jul. 2020, pp. 187–196.

[33] J. F. DeRose, J. Wang, and M. Berger, "Attention flows: Analyzing and comparing attention mechanisms in language models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1160–1170, 2021.

[34] Z. J. Wang, R. Turko, and D. H. Chau, "Dodrio: Exploring Transformer Models with Interactive Visualization," in *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2021, pp. 132–141.

[35] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? an analysis of BERT's attention," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 276–286.

[36] S. Jain and B. C. Wallace, "Attention is not Explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3543–3556.

[37] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "A diagnostic study of explainability techniques for text classification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3256–3274.

[38] H. Strobelt, A. Webson, V. Sanh, B. Hoover, J. Beyer, H. Pfister, and A. M. Rush, "Interactive and visual prompt engineering for ad-hoc task adaptation with large language models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 1146–1156, 2023.

[39] H. Strobelt, B. Hoover, A. Satyanaryan, and S. Gehrmann, "LMdiff: A visual diff tool to compare language models," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 96–105.

[40] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[41] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, "Automatic labelling of topic models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. USA: Association for Computational Linguistics, 2011, p. 1536–1545.

[42] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ser. ACL '94. USA: Association for Computational Linguistics, 1994, p. 133–138.

[43] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[44] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.

[45] J. H. W. Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[46] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[47] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[48] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, p. 39–41, nov 1995.

[49] B. B. Bederson, "Fisheye menus," in *Proceedings of the 13th annual ACM symposium on User interface software and technology*, 2000, pp. 217–225.

[50] K. A. Olsen, R. R. Korfhage, K. M. Sochats, M. B. Spring, and J. G. Williams, "Visualization of a document collection: The vibe system," *Information Processing & Management*, vol. 29, no. 1, pp. 69–81, 1993.

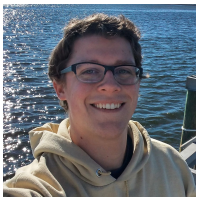
[51] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "PubMedQA: A dataset for biomedical research question answering," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2567–2577.

[52] D. Nozza, F. Bianchi, and D. Hovy, ""HONEST: Measuring hurtful sentence completion in language models"," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 2398–2406.

[53] D. Nozza, F. Bianchi, A. Lauscher, and D. Hovy, "Measuring harmful sentence completion in language models for LGBTQIA+ individuals," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 26–34.

[54] J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K.-W. Chang, and R. Gupta, "Bold: Dataset and metrics for measuring biases in open-ended language generation," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 862–872.

[55] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics," *The University of Chicago Legal Forum*, vol. 1989, no. 1, pp. 139–167, 1989.



Adam Coscia is a PhD student at Georgia Tech's School of Interactive Computing and a member of the Visual Analytics Lab. His research interests include Visual Analytics, Human-Computer Interaction, and Explainable Artificial Intelligence (AI) with Large Language Models and Knowledge Graphs. He received his B.S. in Physics. He won the President's Fellowship for top incoming PhD students.

Alex Endert is an associate professor at the School of Interactive Computing, Georgia Tech. He directs the Visual Analytics Lab, which explores novel user interaction techniques for visual analytics.