Evaluating Readability and Faithfulness of Concept-based Explanations

Meng Li[1*]**, Haoran Jin**[2*]**, Ruixuan Huang**[2]**, Zhihao Xu**[1]**,**
Defu Lian[2]**, Zijia Lin**[3]**, Di Zhang**[3]**, Xiting Wang**[1†]
[1] Renmin University of China
[2] University of Science and Technology of China [3] Kuaishou Technology

Abstract

With the growing popularity of general-purpose Large Language Models (LLMs), comes a need for more global explanations of model behaviors. Concept-based explanations arise as a promising avenue for explaining high-level patterns learned by LLMs. Yet their evaluation poses unique challenges, especially due to their non-local nature and high dimensional representation in a model's hidden space. Current methods approach concepts from different perspectives, lacking a unified formalization. This makes evaluating the core measures of concepts, namely faithfulness or readability, challenging. To bridge the gap, we introduce a formal definition of concepts generalizing to diverse concept-based explanations' settings. Based on this, we quantify the faithfulness of a concept explanation via perturbation. We ensure adequate perturbation in the high-dimensional space for different concepts via an optimization problem. Readability is approximated via an automatic and deterministic measure, quantifying the coherence of patterns that maximally activate a concept while aligning with human understanding. Finally, based on measurement theory, we apply a meta-evaluation method for evaluating these measures, generalizable to other types of explanations or tasks as well. Extensive experimental analysis has been conducted to inform the selection of explanation evaluation measures. [1]

1 Introduction

Explainable Artificial Intelligence (XAI) holds significant value in pre-trained language models' mechanism understanding (Li et al., 2022), visualization (Yang et al., 2024), performance enhancement (Wu et al., 2023b; Ribeiro et al., 2016; Wang et al., 2022), and security (Burger et al., 2023;

Zou et al., 2023). Previous XAI algorithms have been applied to NLP tasks (Wu et al., 2023a), vision tasks (Wang et al., 2023b) and recommendation (Jin et al., 2022; Yang et al., 2022). These include natural language explanation (Zhang et al., 2024; Lee et al., 2022), attention explanation (Chen et al., 2019b; Gao et al., 2019), and especially attribution methods (Lundberg and Lee, 2017; Sundararajan et al., 2017; Guan et al., 2019). The attribution methods identify "where" the model looks rather than "what" it comprehends (Colin et al., 2022), typically offering local explanations for a limited number of input samples, restricting their utility in practical settings (Colin et al., 2022; Adebayo et al., 2018). **Concept**-based explanations (Kim et al., 2018; Cunningham et al., 2023; Fel et al., 2023b) can mitigate the limitations of attribution methods by recognizing high-level (Kim et al., 2018) patterns (see Fig. 1), which provide concise, human-understandable explanations of models' internal state.

Despite these merits, the development of concept-based explanations may be hindered due to a lack of standardized and rigorous evaluation methodology. Unlike a single importance score assigned on each scalar input by attribution methods, diverse explanation methods approach high-dimensional concepts from different aspects. This includes a single classification plane (Kim et al., 2018), an overcomplete set of basis (Cunningham et al., 2023), or a module designed beforehand (Koh et al., 2020), lacking a unified landscape (*C1*). Moreover, its non-local nature across samples (Kim et al., 2018), combined with the high cost of human evaluation when the number of concepts is large, makes evaluating a concept's readability challenging (*C2*). For available evaluation measures (Hoffman et al., 2018), it is difficult to test their reliability and validity (*C3*).

In this paper, we address the challenges above and make the following contributions:

---

[*]These authors contributed equally to this work.
[†]Corresponding author: xitingwang@ruc.edu.cn
[1]Codes available at https://github.com/hr-jin/Concept-Explanation-Evaluation
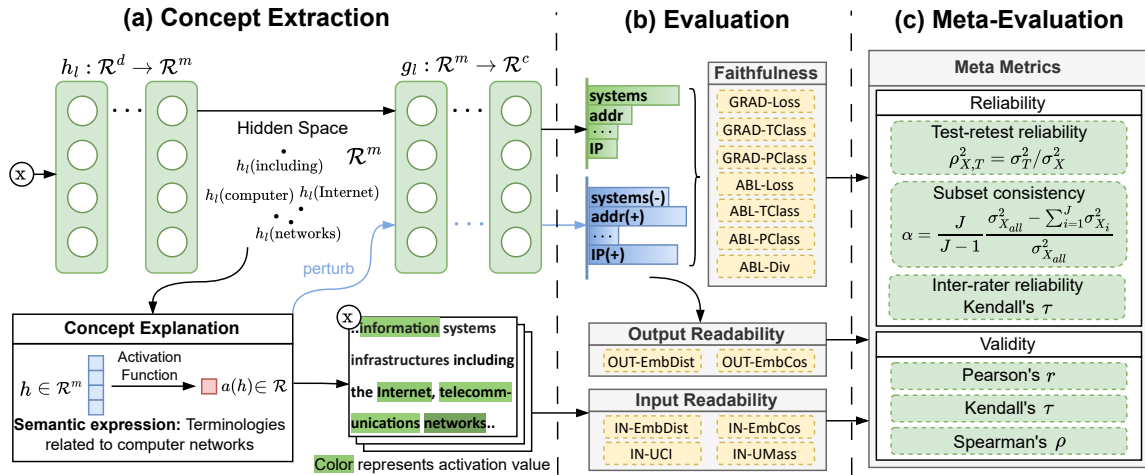
Figure 1: The overall framework. (a) Concept extraction: We formalize concepts as virtual neurons. (b) Evaluation is approached via readability and faithfulness. Readability is approximated by the semantic similarity of patterns that maximally activate the concept. Faithfulness is approximated by the difference in output when a concept is perturbed. (c) Meta-Evaluation is performed on the observed results of proposed measures via reliability and validity.

First, we **provide a unified definition of diverse concept-based explanation methods and quantify faithfulness under this formalization** (*C1*). By summarizing common patterns of concept-based explanation, we provide a formal definition of a concept, which can generalize to both supervised and unsupervised, post-hoc and interpretable-by-design methods, language and vision domains. Based on this, we quantify the faithfulness of a concept explanation via perturbation. We ensure adequate perturbation in the high-dimensional space for different concepts via an optimization problem.

Second, we **approximate readability via coherence of patterns that maximally activates a concept** (*C2*). We utilize the formulation defined above to recognize patterns across samples that maximally activate a concept, from both the input and the output side. Then, we estimate how coherent they are as one concept via semantic similarity. Experimental results have shown this automatic measure correlates highly with human evaluation.

Third, we **apply the classic measurement theory to perform a meta-evaluation on the faithfulness and readability measures** (*C3*). Measurement theory (Allen and Yen, 2001; Xiao et al., 2023) has been long utilized to verify whether a measurement is reliable and valid. Approaching via reliability and validity, this meta-evaluation method is useful for evaluating the measures for concepts and can be generalized to analyze the effectiveness of other measures, for example, measures for other types of explanations and other natural language tasks. Experimental results have filtered out 4 measures with low reliability, i.e. *LLM-Score,*

GRAD-Loss, IN-UCI, IN-UMass, and verified the remaining faithfulness and readability measures' validity.

2 Concept Formalization

In this paper, we primarily focus on explaining LLMs as black-box models. Meanwhile, our method can be generalized to many other deep classification models, including image models (see Appx.C). As illustrated in Fig. 1, we consider the black-box model to take an input x from a dataset D and output y, a k-class classification result. In text generation, k is the vocabulary size. For the l-th layer to be interpreted, given a sequence of input tokens $x^1, ..., x^t$, their corresponding hidden representations are $h_l^1, ..., h_l^t$. The output classification logits are $g(h)$.

Within the context of Deep Neural Networks (DNNs), we summarize common patterns of concepts and establish a unified framework. Specifically, each concept is represented as a virtual neuron defined by an activation function that maps a hidden representation h into a real value $a : \mathcal{R}^m \to \mathcal{R}$, where a positive output signifies activation. For each concept, a semantic expression may be given by humans or LLMs, depending on the concept explanation methods (Kim et al., 2018; Bills et al., 2023). Some methods take concepts and semantic expressions predefined by humans as inputs (e.g., (Kim et al., 2018)), while others require additional steps to produce a semantic expression based on highly activated tokens and samples of the extracted concepts (Bills et al., 2023). Specifically, given high-activation samples of the concept and

the highly activated tokens in these samples (e.g., "Internet, computer, networks, . . . "), an LLM or a human labeler provides a semantic expression that summarizes their common patterns (e.g., "Terminologies related to computer networks") (Bills et al., 2023).

Our formalization can integrate diverse concept explanation methods, as shown in Tab. 1. This includes both supervised methods that require prior information about concepts (e.g., input samples that contain and do not contain the concepts) (Kim et al., 2018) and unsupervised methods that do not rely on such prior information (Ghorbani et al., 2019). Our method also works for both post-hoc explanation methods that interpret a model after it is trained (Kim et al., 2018) and interpretable-by-design approaches that integrate interpretability mechanisms directly into the model's architecture before training (Koh et al., 2020). Additionally, it applies to image backbone models as well.

3 Concept Evaluation Measures

We have conducted a literature survey on evaluation measures for concept-based explanations (Fig. 5 in Appx. A), and decided to focus on two aspects that are of common interest: testing how well they reflect the underlying mechanisms of the machine (**faithfulness**) and assessing the extent to which explanations can be understood by humans (**readability**).

3.1 Faithfulness

Widely studied in previous XAI methods, faithfulness is crucial for assessing how well a concept reflects a model's internal mechanism (Chan et al., 2022; Lee et al., 2023; McCarthy and Prince, 1995). However, its direct application to concept-based explanations presents challenges, particularly due to concepts' ambiguous representation in the hidden space of a model. The adequate degree of perturbation needed for diverse concepts extracted may vary, making it difficult to ensure a fair comparison.

We quantify the faithfulness of a concept by the change in the output $g(h)$ after perturbing the hidden representation h in the hidden space $\mathcal{H}$ where the concepts reside. We formulate faithfulness as $\gamma(a, \xi, \delta)$, where $\xi(h, a)$ applies a perturbation on h given the activation function $a(h)$, and $\delta(y, y')$ measures the output difference.

$$\gamma(a, \xi, \delta) = \frac{1}{|x|} \sum_{h^t \in f(x)} \delta(y, y') \quad (1)$$

with $y = g(h^t), y' = g(\xi(h^t, a))$ being the probability distribution of output vocabulary.

Concept perturbation. Based on the formalization of concepts in Sec. 2, we view this problem as an optimization problem. As the concept formalization provided above encapsulates diverse kinds of concepts, this transformation allows the perturbation strategies to generalize beyond the linear form of concepts, like (Chen et al., 2019a).

Typical perturbation strategies include: 1) ξ_e: concept ϵ-addition, wherein a near zero ϵ is introduced to maximally increase concept activation; 2) ξ_a: concept ablation, which involves removing all information of the concept. The optimization problems can be formulated as:

$$\xi_e(h, a) = \arg\max_{h'} a(h'), \quad \text{s.t. } |h' - h| = \epsilon \quad (2)$$

$$\xi_a(h, a) = \arg\min_{h'} ||h' - h||_2^2, \text{ s.t. } a(h) = 0 \quad (3)$$

When the activation function is linear, e.g., $a(h) = v^T h$, the above problems have closed-form solutions (detailed derivation in Appx. B): Correspondingly, the two perturbation strategies are:

$$\text{(GRAD)} \quad \xi_e(h, a) = \lim_{\epsilon \to 0} h + \epsilon v \quad (4)$$

$$\text{(ABL)} \quad \xi_a(h, a) = h - \frac{v^T h}{v^T v} v \quad (5)$$

Output difference. To quantify different aspects of faithfulness, we include i) difference in training loss (δ_l), ii) deviation in logit statistics (δ_h), iii) difference in the logit prediction of class j (δ_c),

$$\text{(Loss)} \quad \delta_l(y, y') = \mathcal{L}(y, y^*) - \mathcal{L}(y', y^*) \quad (6)$$

$$\text{(Div)} \quad \delta_h(y, y') = H(y, y') \quad (7)$$

$$\text{(Class)} \quad \delta_c^j(y, y') = -(y^j - y'^j) \quad (8)$$

Here, $\mathcal{L}$ is a certain loss function (Schwab and Karlen, 2019; Bricken et al., 2023), y, y' are the output classfication logits, y^* is corresponding ground truth label, y^j, y'^j are the logits of class j. To quantify the discrepancy between distributions, we utilize a statistic H, specifically KL-Divergence in our experimental setup.

For ease of reference, perturbations are expressed as prefixes, and difference measures are denoted as suffixes. Furthermore, we divide *Class* into *PClass* (prediction class) and *TClass* (true class) with j taking the predicted token class or ground truth token class. For instance, faithfulness computed via gradient to prediction class, as

	Method	Modal	Activation function $a(h)$
Supervised	TCAV (Kim et al., 2018)	text/image	$v^T h + b$
	CBM* (Koh et al., 2020)	image	$o_i^T h$
	ProtoPNet* (Chen et al., 2019a)	image	$\max_{\tilde{h}\in\text{patches}(h)} \dfrac{\log((\|\|\tilde{h} - v\|\|_2^2 + 1)}{\|\|\tilde{h} - v\|\|_2^2 + \epsilon}$
Unsupervised	NetDissect (Bau et al., 2017)	image	$\dfrac{M(h) \cap L_c(x)}{M(h) \cup L_c(x)}$
	Neuron (Bills et al., 2023)	text/image	$o_i^T h$
	SAE (Cunningham et al., 2023)	text	$\text{ReLU}(v^T h + b)$

Table 1: Concept-based explanations' activation function. * denotes interpretable-by-design methods. Hyperparameters: 1) v, o is a concept vector within the same space as h, and o_i denotes a one-hot vector where i indicates the position of the 1 in the vector. 2) $M(h)$ selects the top-quantile activations and upsample them to the same dimension as x, and $L_c(x)$ is a pixel-level human-annotated label on x. 3) b is a bias term.

proposed in (Kim et al., 2018), is represented as *GRAD-PClass*. Altogether, there are 2*4 kinds of available faithfulness measures. As the gradient option is too slow on vectors, we leave out *GRAD-Div*.

3.2 Readability

Readability assesses the extent to which humans can comprehend the extracted concept (Lage et al., 2019). Most of the time, when patterns that maximally activate a concept are coherent (see example in Fig. 1), can the concept be easily understandable to humans. We design coherence measures based on OpenAI's pipeline (Bills et al., 2023) for human evaluation of concept quality. They presented human labelers with fragments where highly activated tokens were shown with color highlighting and asked the humans to try summarizing the commonalities of these highly activated tokens. We automate this process by assessing the commonality of highly activated tokens via co-occurrence or embedding similarity.

As cross-sample patterns are extracted from a large corpus, diverse samples are needed to evaluate a concept's readability. Although previous efforts have made some progress in evaluating readability, they confront the challenge of ensuring data comprehensiveness while minimizing cost. Tab. 2 compares different measures for readability, including human evaluation (Kim et al., 2018; Ghorbani et al., 2019), LLM-based measures (Bills et al., 2023; Singh et al., 2023), and our proposed coherence-based measures. For the LLM-based evaluation, we considered (Bills et al., 2023; Bricken et al., 2023), which used less than 100 samples. For human evaluation, we considered the

classical method by (Ghorbani et al., 2019), where each human rater scored no more than 20 samples per concept.

Method	#Sample	Cost	Reliability
Human	< 20	high	medium
LLM-based	< 100	medium	low
Ours	> 2000	low	high

Table 2: Comparison of readability measures. #Sample denotes the maximum number of samples applicable for evaluating a concept.

Human evaluation. Existing approaches predominantly rely on case studies and user studies (Kim et al., 2018; Ghorbani et al., 2019; Chen et al., 2019a), asking humans to score a concept given a limited number of demonstrative samples. They are subject to issues of validation, standardization, and reproducibility (Clark et al., 2021; Howcroft et al., 2020).

LLM-based. As inexpensive human substitutes, LLMs have been utilized in evaluating concept-based explanations. A typical LLM-based score (Bills et al., 2023; Singh et al., 2023) is obtained by: 1) letting LLM summarize a natural language explanation s for the concept (e.g., semantic expression in Fig. 1) given formatted samples that maximally activates on the concept and activations a; 2) letting LLM guess the activation given only sample text and the generated explanation; 3) calculating an explanation score based on the variance between true activation and the simulated activation. However, the number of samples inputted to LLMs (4 in (Bills et al., 2023)) in step 1 is limited to maximum input length. This limits the comprehensiveness of the generated explanation, as shown in a case study in Appx. D. Even if the maximum

input length is extended to 200k+ like Claude 3[2], it may suffer from high computation cost and poor performance in long-dependency tasks (Li et al., 2023).

Coherence-based. To address these limitations, we propose novel measures inspired by topic coherence. Topic coherence measures are widely used in the field of topic modeling to estimate whether a topic identified from a large corpus can be easily understood by humans (Newman et al., 2010). Here, the basic idea is to approximate readability based on the semantic similarity between patterns that maximally activate a concept: we estimate how coherent they are as one topic (Fig. 1). These measures mainly rely on the concept activation function, allowing for scalable, automatic, and deterministic evaluation.

Patterns that maximally activate a concept are obtained as follows. Initially, a subset of texts is selected and processed through a black-box LLM to obtain concept-specific activations for each token. High-activation tokens, indicative of a strong association with the analyzed concept, are then identified. For these tokens, important contextual words are extracted by ablating each word in the context and identifying those that impose the most impact on the high-activation token. Similar information can be obtained from the output side. We extract tokens with the top-k highest likelihood when setting the hidden representation highly active on the concept and not on others.

For our evaluation, we employ semantic similarity measures including *UCI* (Newman et al., 2009), *UMass* (Mimno et al., 2011), and two deep measures *Embedding Distance* (*EmbDist*), *Embedding Cosine Similarity* (*EmbCos*). Each measure computes similarity $\mu(x^i, x^j)$ between two tokens x^i, x^j as follows:

$$\mu_{\text{UCI}}(x^i, x^j) = \log \frac{P(x^i, x^j) + \epsilon}{P(x^i)P(x^j)} \quad (9)$$

$$\mu_{\text{UMass}}(x^i, x^j) = \log \frac{P(x^i, x^j) + \epsilon}{P(x^j)} \quad (10)$$

$$\mu_{\text{EmbDist}}(x^i, x^j) = - \|e(x^i) - e(x^j)\|_2 \quad (11)$$

$$\mu_{\text{EmbCos}}(x^i, x^j) = \frac{e(x^i) \cdot e(x^j)}{|e(x^i)||e(x^j)|} \quad (12)$$

Probabilities are estimated based on word occurrence frequency in the corpus. To prevent zero values in logarithmic operations, a small value ϵ

is introduced. $e(x^i)$ embeds a word to a continuous semantic space, for example, using embedding models like BERT.

For ease of reference and consistency, we denote readability on the input/output side using the prefixes *IN/OUT*. For instance, readability computed using *UCI* similarity on the input side is represented as *IN-UCI*. Note that coherence-based measures may not capture all the desiderata of a readable explanation. Yet, it is still of interest to utilize this measure to filter a large amount of concepts when human evaluation may not be applicable.

4 Meta Evaluation

How can we discern the effectiveness among possible measures available for evaluating concept-based explanations? Borrowing metrics from measurement theory (Allen and Yen, 2001) and psychometrics (Wang et al., 2023c; Xiao et al., 2023), our meta-evaluation focus centers on **reliability** and **validity**, guided by the methodological framework outlined in (Allen and Yen, 2001). Our meta-evaluation methods can generalized to measures of a broader scope, including other XAI methods and other natural language tasks like generation.

4.1 Reliability

Reliability is crucial for assessing the consistency of a measure under multiple measurements, accounting for random errors introduced during measurement. These errors can arise from non-deterministic algorithms, data subsets, and human subjectivity. We particularly focus on three aspects: 1) **test-retest reliability**, quantifying the expected amount of uncertainty in the observed measure 2) **subset consistency**, measured as fluctuation across data subsets within a test; 3) **inter-rater reliability**, quantifying the degree of agreement between two or more raters.

Test-retest reliability is quantified as the test-retest correlation: on the concepts extracted, we compute the same measure twice for each concept. The Pearson correlation (Galton, 1877) between the two sets of results is test-retest reliability, which is an estimate of the expectation of:

$$\rho_{X,T}^2 = \frac{\sigma_T^2}{\sigma_X^2} \quad (13)$$

where X is the observed score, and T is the true score, σ_*^2 denotes the variance of a random variable $*$. Typically, the minimal standard for an ac-

ceptable measure is 0.9 (Nunnally and Bernstein, 1994).

Subset consistency is estimated through Cronbach's Alpha (Cronbach, 1951), a classic coefficient for evaluating internal consistency in measurement theory:

$$\alpha = \frac{J}{J-1} \frac{\sigma^2_{X_{all}} - \sum_{j=1}^{J} \sigma^2_{X_j}}{\sigma^2_{X_{all}}} \quad (14)$$

$X_1, X_2, ... X_J$ are results of measure X across different data subsets. The overall score on the entire dataset is expressed as $X_{all} = \sum_{j=1}^{J} X_j$. α is the lower bound of squared correlation $\rho^2_{X,T}$ of observed score X and true score T (Cronbach, 1951). For a measure with low subset consistency, one may use a larger test dataset to ensure the result's consistency.

Inter-rater reliability measures the degree of agreement across raters, calculated as score correlation among them. In this paper, we apply Kendall's τ (Kendall, 1938) to measure pairwise correlation among raters using a scale that is ordered:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}(X_i^1 - X_j^1)\text{sgn}(X_i^2 - X_j^2) \quad (15)$$

X_i^* denotes the score on the i-th concept given by rater $*$. Evaluations that rely on humans must exhibit good inter-rater reliability, or, they are not reliable tests.

4.2 Validity

Validity is crucial in assessing how well a test measures the intended construct (Nunnally and Bernstein, 1994). A construct refers to the underlying criterion to be measured. In our case, it is faithfulness or readability. We focus on **concurrent validity**, evaluating the extent to which a test score predicts outcomes on a validated measure (Cronbach and Meehl, 1955), and **construct validity**, examining how well indicators represent an unmeasurable concept (Cronbach and Meehl, 1955). Construct validity can be further divided into convergent validity and divergent validity.

Concurrent validity reflects the appropriateness of a measure as an alternative to an existing reference, quantified via the correlation between the two scores. For example, an automatic measure for readability is used to approximate human evaluation at a large scale. Only when the automatic measure for readability is highly correlated with human scores, can we treat it as an approximate of human evaluation. Here we use classical correlation metrics to estimate concurrent validity (Kendall, 1938; Spearman, 1961; Galton, 1877). Note that random error in either the automatic measure or human evaluation may impair concurrent validity. Thus being reliable is a premise of being valid.

Convergent validity verifies whether measures of the same construct are indeed related. For example, whether the purposed faithfulness measures are related to each other. As the underlying construct is often inaccessible to directly assess the measures' concurrent validity, convergent validity provides a statistic tool to assess construct validity via its relation (Kendall, 1938) with other measures of the same construct.

Divergent validity tests whether measures of unrelated constructs are indeed unrelated. For example, for distinct aspects considered of concept-based explanation (e.g., readability and faithfulness), measures of different aspects should show a significantly lower correlation than measures of the same aspect. Here we apply Kendall's τ (Kendall, 1938) as a measure of correlation. A bad divergent validity may indicate potential bias in designed measures, calling for a more rigorous inspection of potential bias.

To inspect the construct validity of the measures to the intended constructs, we employ the multitrait-multimethod (MTMM) table methodology introduced by (Campbell and Fiske, 1959). This table conventionally presents pairwise correlations of observed measure scores on the off-diagonals and the subset consistency of each score on the diagonals.

5 Experiments

5.1 Datasets and Experimental Settings

We leverage the Pile dataset, a comprehensive collection curated by (Gao et al., 2020), which stands as the largest publicly available dataset for pre-training language models like Pythia (Biderman et al., 2023). This dataset includes a vast 825 GiB of diverse data and encompasses 22 smaller, high-quality datasets spanning multilingual text and code. Its rich diversity facilitates the extraction of a wide array of concepts, crucial for our evaluation framework.

For the backbone model, we choose Pythia due to its pre-training on the Pile dataset, ensuring consistent knowledge representation between the training and explanation phases. Additionally, we

include GPT-2 (Radford et al., 2019) to ensure the consistency of our findings across backbones (Appx. E). Further details on these models are provided in Tab. 6. To eliminate the impact of random fluctuations, we test each measure across 10 batches, each comprising 256 sentences with 128 tokens, totaling 327,680 tokens.

5.2 Comparison of Evaluation Measures

In this section, we evaluate our proposed concept-based explanation measures, employing the meta-evaluation method for thorough assessment. To ensure a fair comparison, we randomly sampled 100 concepts extracted by each unsupervised baseline applicable to the language domain on the same backbone model, including Neuron-based method (Bills et al., 2023) and Sparse Autoencoder (Cunningham et al., 2023). We primarily introduce results from the middle layer of Pythia-70M, with other consistent results across different layers and models in Appx. E. Due to the possibility of highly enhanced tokens not appearing in the dataset, we apply *UCI* and *UMass* measures only on the input side.

5.2.1 Reliability

In this section, we analyze which measures are reliable to random noise introduced by retesting, different data subsets, and human subjectivity.

Test-retest reliability results, depicted in Fig. 2, verifies the deterministic nature of the proposed measures, except for *LLM-Score* (Bills et al., 2023). *LLM-Score* is less acceptable, which may be due to the inherent randomness introduced by sampling the most probable tokens.
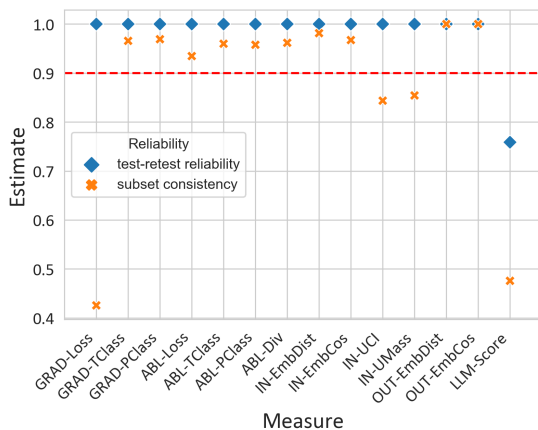
Figure 2: Estimated test-retest reliability and subset consistency of the proposed measures. The red dashed line indicates the minimal standard of 0.9 (Nunnally and Bernstein, 1994).

Subset consistency provides further filtering of

present measures with a threshold of 0.9 (Nunnally and Bernstein, 1994), as shown in Fig. 2. For the faithfulness family, *GRAD-Loss* shows an undesirable low consistency, probably due to the coupling of gradient and loss during training. For the readability family, *IN-UCI* and *IN-Umass* is less acceptable, attributing to the diverse nature of different concept's n-grams. Moreover, their capability to capture semantic similarity is also less desirable according to a case study shown in Appx. D

Inter-rater reliability is tested on human evaluation of readability. The concepts used for analysis above are scored by each human labeler with a high school level of English proficiency. They are blinded to the source method for the generated concepts and are tasked with scoring each concept on a scale of 1 to 5 based on two criteria: input readability and output readability. The recruitment of the experts and the setting of the user study are detailed in Appx. G.

	Input	Output	Average
Expert1 & Expert2	0.81	0.77	0.79
Expert1 & Expert3	0.76	0.75	0.76
Expert2 & Expert3	0.74	0.72	0.73

Table 3: Experts' Kendall's τ correlation as inter-rater reliability.

Tab. 3 shows the inter-rater reliability. Overall, experts' correlations are high, with an average of 0.77 and 0.75 on the input and output sides.

5.2.2 Validity

Here, we analyze whether the measures assess the intended construct, i.e., readability or faithfulness. We leave out *LLM-Score*, *GRAD-Loss*, *IN-UCI*, *IN-UMass* due to their low reliability as discovered in Sec. 5.2.1.

	Kendall		Pearson		Spearman	
	IR	OR	IR	OR	IR	OR
LLM-Score	<u>0.54</u>	0.09	**0.70**	0.12	<u>0.67</u>	0.12
IN-EmbDist	0.19	0.12	0.27	0.16	0.26	0.16
IN-EmbCos	**0.56**	0.18	<u>0.68</u>	0.18	**0.70**	0.24
OUT-EmbDist	0.15	<u>0.63</u>	0.16	<u>0.73</u>	0.21	<u>0.76</u>
OUT-EmbCos	0.17	**0.67**	0.16	**0.75**	0.23	**0.80**

Table 4: Concurrent validity of Input Readability (IR) and Output Readability (OR). The best results are marked in **bold**. The second-best results are <u>underlined</u>.

Concurrent validity. In this experiment, we treat the user study results for readability as a criterion measure. Tab. 4 shows how well existing automatic measures for readability correlate with user

study results. *IN-EmbCos* is the top-performing measure to predict input readability (IR), and *OUT-EmbCos* is the best in predicting output readability (OR). This demonstrates the effectiveness of our coherence-based measure EmbCos as an approximation of human evaluation. Compared with LLM-based measure that requires expensive API calls to GPT-4, EmbCos has a stronger correlation with human labels while requiring a much smaller computational cost. We recommend EmbCos as an inexpensive substitute for human evaluation, especially on large-scale evaluations. Yet human evaluation is still needed for more rigorous analysis.

In Fig. 3, the off-diagonals visually demonstrate the **construct validity** between our proposed measures. Our observations are as follows.
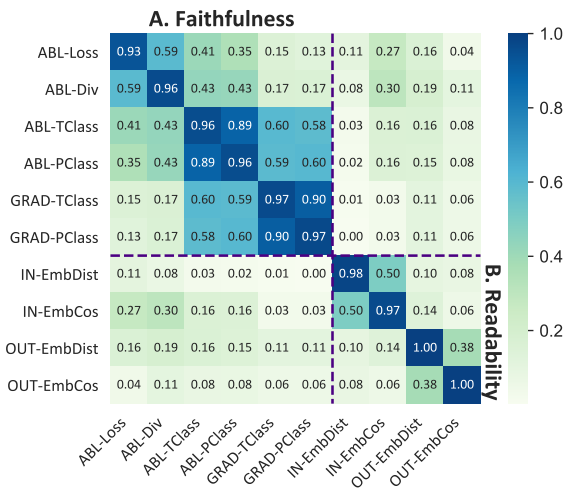

Figure 3: The MTMM table of the evaluation measures: 1) subset consistency is shown on the diagonals; 2) construct validity is displayed on the off-diagonals.

Divergent Validity is inspected via correlation between unrelated measures. Measures of faithfulness (A) show a low correlation (0.0-0.3) with measures of readability (B), revealing their distinct nature, which is as expected. Input readability and output readability are also divergent (correlation less than 0.15), demonstrating concepts' unique patterns on both sides. While previous efforts on readability mostly focus on the input side, more careful inspection on the output side is needed.

Convergent Validity is inspected via correlation between measures of the same construct. Faithfulness measures (A) displayed moderate correlations in general, averaging around 0.5. Agreement between measures with the same perturbation strategy or difference measurement is higher than others, indicating their potential relation. *-TClass* and

-PClass showed a higher correlation, due to the consistency between prediction and true classes in well-trained language models. In the meantime, the agreement of readability measures (B) on either the input side or output side is moderate.

Our findings are consistent across different layers and backbones. Interested readers may refer to Appx. E for detailed results.

5.3 Comparison of Explanation Methods

We conducted a comparative assessment of three different baseline methods on the language domain, including the concepts of neuron (Bills et al., 2023), sparse autoencoder (Cunningham et al., 2023), and TCAV (Kim et al., 2018). The results for both the neuron and sparse autoencoder were computed as the average values across 100 randomly sampled concepts from the concept set. We derive the supervised concept using TCAV following (Kim et al., 2018; Xu et al., 2024). Initially, LLM's harmful QA (Bhardwaj and Poria, 2023) is treated as positive examples, and random texts are treated as negative examples. Their hidden representations are used to train a linear classifier, which aims to differentiate the representations of positive examples from negative ones. The trained classifier is treated as the concept's activation function. The results of this analysis are shown in Fig. 4.
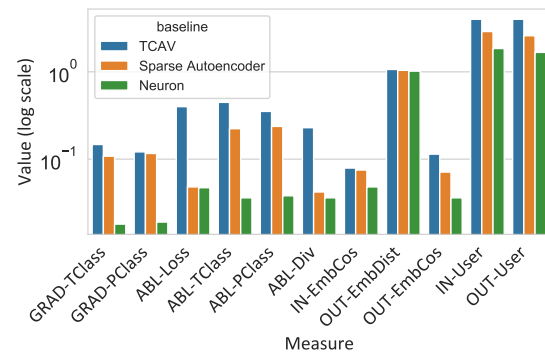
Figure 4: Performance of different baselines on representative measures.

Sparse autoencoder surpasses the neuron-based methods across all evaluated measures, which is as expected. Nevertheless, as an unsupervised method, it falls short of TCAV on these same measures. This implies that the average quality of the concepts it extracted is not as high as the concepts derived from supervised counterparts. Additionally, the discrepancy between human ratings for different baseline methods is smaller than that between other readability measures. Upon detailed

Method	Input Relevant Tokens	Output Preferred Tokens
TCAV	·information, ·sensitive, ·fraudulent, ·purposes, ·violence, information, ·candidate, ·someone, ·stealing, ·hatred	·assassination, ·illegal, ·gren,·rape, ·unconstitutional, impeachment, ·/., ·prosecution, ·unlawful, ·conspiracy
Sparse Autoencoder	·north, ·west, ·east, ·South, ·North, South, ·northern, ·southern, ·eastern, ·dorsal	western, ward, bound, side, ampton, wards, ·facing, line, most, ·coast
	·task, ·carbohydrates, ·radiation, ·musician, front, ·version, own, ·control, ·Hope, ·caution	·answer, ·tumor, ·disambiguation, któ, omitempty, ·Version, ·World, ·stream, ·huh, ·UK
Neuron	·gap, ·als, ·going, ·3, ·mit, ·maybe, ·True, ·t, ·c, ·URN	lement, ters, right,uki, ter, ecycle, aut, ·β, er, ·\n\n

Table 5: Patterns that maximally activate some demonstrative concepts of the baselines. '·' indicates space. For sparse autoencoder, we selected one concept from both the top 10% and bottom 10% based on the average rank results of *IN-EmbCos* and *OUT-EmbCos*. For the neuron method, we only showcased the top concepts.

analysis of the results, it appears that human raters tend to give less discriminative scores ranging from 2 to 4, rarely awarding a 1 or 5, whereas automated measures show a greater range in scoring.

We also present a case study in Tab. 5 to visually illustrate the readability of concepts extracted by the three baselines. Firstly, TCAV's extracted concept shows high readability, with both input and output key tokens strongly tied to the "harmful QA" training theme. Secondly, The performance of the sparse autoencoder is notably inconsistent, whose concept set varies widely in readability measures. However, on average, upon observing many concepts, we found that the readability of concepts extracted by sparse autoencoder surpasses that of neurons. This suggests that the sparse dictionary paradigm generally enhances the quality of the entire concept set, mitigating the issue of superposition (Elhage et al., 2022).

Besides, we found that LLM has learned a seemingly redundant yet interesting pattern for the first concept shown for sparse autoencoder (e.g., north,·west, ·east, ·South, ·North, South, ·northern). Though these tokens are quite similar for humans, we do not know whether they are considered the same for LLMs. The embedding similarity between these tokens reflects LLMs' ability to model them just like how humans perceive them as similar.

6 Conclusion

This paper introduced two automatic evaluation measures, readability and faithfulness, for concept-based explanations. We first formalize a general definition of concepts and quantify faithfulness under this formalization. Then, we approximate readability via the coherence of patterns that maximally activate a concept. Another key contribution of this paper is that we describe a meta-evaluation method for evaluating the reliability and validity of these evaluation measures across diverse settings based on measurement theory. Through extensive experimental analysis, we inform the selection of explanation evaluation measures, hoping to advance the field of concept-based explanation.

Limitations

Our framework may not encompass the entirety of the concept-based explanation landscape. Although the focus on readability and faithfulness aligns with prior research suggestions (Jacovi and Goldberg, 2020; Lage et al., 2019) and represents core components of evaluating concept-based explanations. We acknowledge that our study represents a modest step towards evaluating concept-based explanations. Future research on other aspects like robustness and stability is necessary.

Topic coherence is not designed to be the ultimate or perfect solution for measuring readability. Other aspects of readability, such as meaningfulness (Ghorbani et al., 2019), may also worth exploring. In the future, we are interested in investigating how these aspects could be quantified automatically, building a more comprehensive landscape of readability.

Due to limited GPU resources and budget constraints, we used smaller versions of LLM, focusing primarily on the 3rd layer of Pythia-70M for our analysis. And our evaluation of the LLM-Score

was restricted to 200 concepts, incurring a cost of around \$1 for a single concept. While this setup, on par with (Cunningham et al., 2023) and more general than (Bricken et al., 2023), allowed for fast analysis and comparison with existing literature, expanding our analysis to larger models could yield more insightful conclusions in the future.

Acknowledgements

Ethical Statements

Our evaluation metrics for concept-based explanations offer a valuable contribution to enhancing human comprehension of LLM. However, it's crucial to acknowledge the potential presence of inherent hallucinations in the evaluation process that may have gone unnoticed.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Mary J Allen and Wendy M Yen. 2001. *Introduction to measurement theory*. Waveland Press.

David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549.

Rishabh Bhardwaj and Soujanya Poria. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Christopher Burger, Lingwei Chen, and Thai Le. 2023. "are your explanations reliable?" investigating the stability of lime in explaining text classifiers by marrying xai and adversarial attack. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12831–12844.

Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81.

Chun Sik Chan, Huanqi Kong, and Guanqing Liang. 2022. A comparative study of faithfulness metrics for model interpretability methods. *arXiv preprint arXiv:2204.05514*.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019a. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Zhi Chen, Yijie Bei, and Cynthia Rudin. 2020. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782.

Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019b. Co-attentive multi-task learning for explainable recommendation. In *IJCAI*, volume 2019, pages 2137–2143.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. 2022. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems*, 35:2832–2845.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2022. Discovering latent concepts learned in bert. *arXiv preprint arXiv:2205.07237*.

Kien Do and Truyen Tran. 2019. Theory and evaluation metrics for learning disentangled representations. *arXiv preprint arXiv:1908.09961*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Mathieu Chalvidal, Thomas Serre, et al. 2023a. A holistic approach to unifying automatic concept extraction and concept importance estimation. *arXiv preprint arXiv:2306.07304*.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. 2023b. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721.

Francis Galton. 1877. Typical laws of heredity 1. *Nature*, 15(388):492–495.

Jingyue Gao, Xiting Wang, Yasha Wang, and Xing Xie. 2019. Explainable recommendation through attentive multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3622–3629.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind W Picard. 2021. Dissect: Disentangled simultaneous explanations via concept traversals. *arXiv preprint arXiv:2105.15164*.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. 2019. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*.

Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a deep and unified understanding of deep neural models in nlp. In *International conference on machine learning*, pages 2454–2463. PMLR.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. *arXiv preprint arXiv:2010.02812*.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. 2022. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5746–5754.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR.

Avinash Kori, Parth Natekar, Ganapathy Krishnamurthi, and Balaji Srinivasan. 2020. Abstracting deep neural networks into concept graphs for concept level interpretability. *arXiv preprint arXiv:2008.06457*.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*.

Jae Hee Lee, Sergio Lanza, and Stefan Wermter. 2023. From neural activations to concepts: A survey on explaining concepts in neural networks. *arXiv preprint arXiv:2310.11884*.

Seungeon Lee, Xiting Wang, Sungwon Han, Xiaoyuan Yi, Xing Xie, and Meeyoung Cha. 2022. Self-explaining deep models with logic rule reasoning. *Advances in Neural Information Processing Systems*, 35:3203–3216.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*.

Zhen Li, Xiting Wang, Weikai Yang, Jing Wu, Zhengyan Zhang, Zhiyuan Liu, Maosong Sun, Hui Zhang, and Shixia Liu. 2022. A unified understanding of deep nlp models for text classification. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4980–4994.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

John J McCarthy and Alan Prince. 1995. Faithfulness and reduplicative identity. *Linguistics Department Faculty Publication Series*, page 10.

Georgii Mikriukov, Gesina Schwalbe, Christian Hellert, and Korinna Bade. 2023. Evaluating the stability of semantic concept representations in cnns for robust explainability. *arXiv preprint arXiv:2304.14864*.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Seil Na, Yo Joong Choe, Dong-Hyun Lee, and Gunhee Kim. 2019. Discovery of natural language concepts in individual units of cnns. *arXiv preprint arXiv:1902.07249*.

David Newman, Sarvnaz Karimi, and Lawrence Cavedon. 2009. External evaluation of topic models. In *Proceedings of the 14th Australasian Document Computing Symposium*, pages 1–8. University of Sydney.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108.

Jum C Nunnally and Ira H Bernstein. 1994. Psychometric theory new york. *NY: McGraw-Hill*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Avi Rosenfeld. 2021. Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th international conference on autonomous agents and multiagent systems*, pages 45–50.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Rafae Khan, and Jia Xu. 2022. Analyzing encoded concepts in transformer language models. *arXiv preprint arXiv:2206.13289*.

Anirban Sarkar, Deepak Vijaykeerthy, Anindya Sarkar, and Vineeth N Balasubramanian. 2022. A framework for learning ante-hoc explainable models via concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10286–10295.

Patrick Schwab and Walter Karlen. 2019. Cxplain: Causal explanations for model interpretation under uncertainty. *Advances in neural information processing systems*, 32.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Chandan Singh, Aliyah R Hsu, Richard Antonello, Shailee Jain, Alexander G Huth, Bin Yu, and Jianfeng Gao. 2023. Explaining black box text modules in natural language with language models. *arXiv preprint arXiv:2305.09863*.

Sanchit Sinha, Mengdi Huai, Jianhui Sun, and Aidong Zhang. 2023. Understanding and enhancing robustness of concept-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15127–15135.

C. Spearman. 1961. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Ao Sun, Pingchuan Ma, Yuanyuan Yuan, and Shuai Wang. 2023. Explain any concept: Segment anything meets concept-based explanation. *arXiv preprint arXiv:2305.10289*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. 2023a. Learning bottleneck concepts in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10962–10971.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, and Tao Kong. 2023b. Densecl: A simple framework for self-supervised dense visual pre-training. *Visual Informatics*, 7(1):30–40.

Xiting Wang, Liming Jiang, Jose Hernandez-Orallo, Luning Sun, David Stillwell, Fang Luo, and Xing Xie. 2023c. Evaluating general-purpose ai with psychometrics. *arXiv preprint arXiv:2310.16379*.

Xiting Wang, Kunpeng Liu, Dongjie Wang, Le Wu, Yanjie Fu, and Xing Xie. 2022. Multi-level recommendation reasoning over knowledge graphs with reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, pages 2098–2108.

Chenwang Wu, Xiting Wang, Defu Lian, Xing Xie, and Enhong Chen. 2023a. A causality inspired framework for model interpretation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2731–2741.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. 2023b. Causal proxy models for concept-based model explanations. In *International conference on machine learning*, pages 37313–37334. PMLR.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing nlg evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982.

Zhihao Xu, Ruixuan Huang, Xiting Wang, Fangzhao Wu, Jing Yao, and Xing Xie. 2024. Uncovering safety risks in open-source llms through concept activation vector. *Advances in Neural Information Processing Systems*.

Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement subgraph reasoning for fake news detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2253–2262.

Weikai Yang, Mengchen Liu, Zheng Wang, and Shixia Liu. 2024. Foundation models meet visualizations: Challenges and opportunities. *Computational Visual Media*, pages 1–26.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565.

Hanyu Zhang, Xiting Wang, Xiang Ao, and Qing He. 2024. Distillation with explanations from large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5018–5028.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. 2021. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11682–11690.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A   Taxonomies

In this section, we present a taxonomy of prior automatic measures for evaluating concept-based explanations based on existing literature on evaluating explainable AI (Hoffman et al., 2018; Jacovi and Goldberg, 2020; Colin et al., 2022). Fig. 5 provides a summarized mind map, offering a visual representation of the various aspects by which concept-based explanation methods can be assessed. We endeavored to use the original terminologies as they appear in the cited works, emphasizing the purposes for which these measures were developed. Due to the evolving nature of the field, some measures might share similarities in their meanings or computational methods, which could lead to perceived overlap.

This makes the selection of suitable evaluation measures hard for practitioners in the field of concept-based explanation Therefore, there is a pressing need for a more unified landscape in the evaluation of concept-based methods to facilitate substantial progress in the field. To address potential confusion, the evaluation measures we propose in this paper seek to clarify and distinguish between the different aspects of evaluation. We aim to pro-
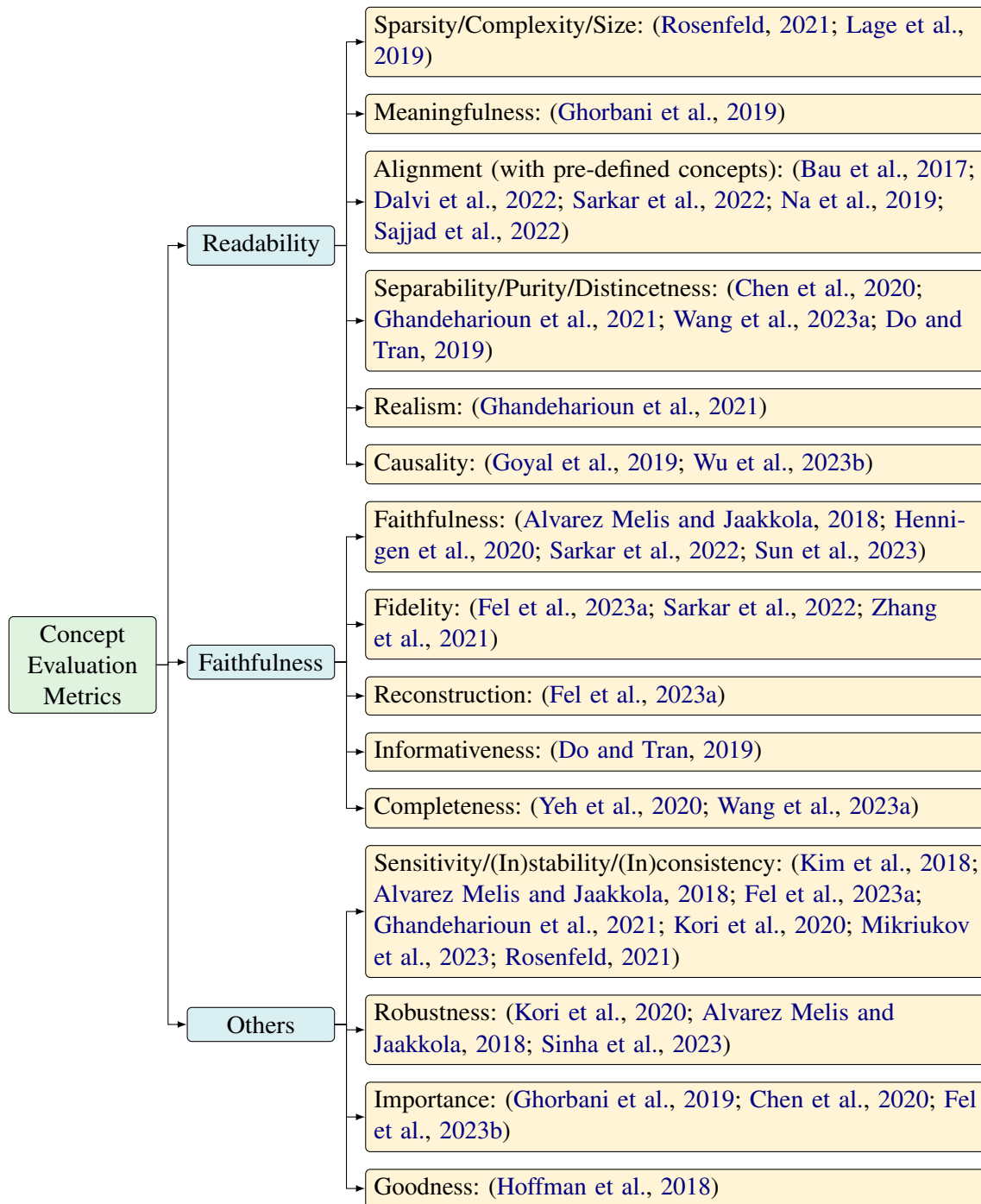
Figure 5: Taxonomy of prior automatic metrics on concept-based explanation methods.

vide a clear and structured approach that reflects the nuanced differences among these measures.

B Derivation of adequate ablation

We consider concept ablation as an optimization problem with a closed-form solution, aiming to minimize perturbation while maintaining zero activation. This optimization problem can be formulated as:

$$\arg\min_{h'} ||h' - h||_2^2, \quad \text{s.t.} \quad \alpha(h) = 0 \qquad (16)$$

We approach this optimization via the Lagrange multiplier. For typical activation function calculated via inner product $\alpha(h) = v^T h$, the Lagrange function is defined as:

$$\mathcal{L}(h, h', v) = ||h' - h||_2^2 + \lambda v^T h' \qquad (17)$$

On stationary points of $\mathcal{L}$:

$$\frac{\delta \mathcal{L}(h, h', v)}{\delta h'} = 0 \qquad (18)$$

$$\Leftrightarrow 2(h' - h) + \lambda v = 0 \qquad (19)$$

$$\Leftrightarrow h' = h - \frac{\lambda}{2} v \qquad (20)$$

As $v^T h' = 0$, we have:

$$v^T(h - \frac{\lambda}{2} v) = 0 \qquad (21)$$

$$\Leftrightarrow \lambda = \frac{2 v^T h}{v^T v} \qquad (22)$$

$$\Leftrightarrow h' = h - \frac{v^T h}{v^T v} v \qquad (23)$$

For disentanglement-based methods, activation is calculated via $\alpha(h) = \text{ReLU}(v^T h + b)$, where $\text{ReLU}(x) = \max(x, 0)$.

$$\text{When} \quad v^T h + b \le 0 \qquad (24)$$

$$\Leftrightarrow \alpha(h) = 0 \qquad (25)$$

$$\Leftrightarrow h' = h \qquad (26)$$

Otherwise, $\alpha(h) = v^T h + b$, the Lagrange function is defined as:

$$\mathcal{L}(h, h', v) = ||h' - h||_2^2 + \lambda(v^T h' + b) \qquad (27)$$

On stationary points of $\mathcal{L}$:

$$\frac{\delta \mathcal{L}(h, h', v)}{\delta h'} = 0 \qquad (28)$$

$$\Leftrightarrow 2(h' - h) + \lambda v = 0 \qquad (29)$$

$$\Leftrightarrow h' = h - \frac{\lambda}{2} v \qquad (30)$$

As $v^T h' + b = 0$, we have:

$$v^T(h - \frac{\lambda}{2} v) + b = 0 \qquad (31)$$

$$\Leftrightarrow \lambda = \frac{2(v^T h + b)}{v^T v} \qquad (32)$$

$$\Leftrightarrow h' = h - \frac{v^T h + b}{v^T v} v \qquad (33)$$

Similarly, we consider concept ϵ-addition as an optimization problem with a closed-form solution, aiming to maximize concept activation with only perturbation of length ϵ. This optimization problem can be formulated as:

$$\arg\max_{h'} a(h'), \quad \text{s.t.} \quad |h' - h| = \epsilon \qquad (34)$$

The solution to this problem when activation function is $a(h) = v^T h$ is:

$$h' = h + \frac{\epsilon}{|v|} v \qquad (35)$$

C Applicability to image domain

In our paper, we mostly focus on LLMs as backbone models. Here we elaborate on how the proposed measures can be extended to the vision domain.

For readability, we can create 'tokens' by adopting a methodology similar to LIME (Ribeiro et al., 2016). Specifically, we can segment each image into superpixels and regard each superpixel as a token in text. These superpixels' embeddings can then be obtained using pre-trained image models like VGG (Simonyan and Zisserman, 2014), and coherence-based measures can be applied by assessing the similarity of these embeddings. While extending measures like *UCI/UMass* to image tasks may present challenges, it remains feasible by first transcribing superpixels into text using vision-language models like CLIP (Radford et al., 2021) and then calculating their co-occurrence. Yet considering the low reliability indicated in Sec. 5.2.1 as well as its original initiative for the language domain, it might be redundant to explore this variant.

Furthermore, faithfulness measures, operating on hidden and output spaces, are inherently independent of data modality and can be directly applied to image tasks. In general, our method can be used as long as a concept can be formulated with a virtual activation function (Sec. 2), which takes a given hidden representation in the model as input and outputs the degree a concept is activated. As

discussed in Sec. 2, we believe this formulation is versatile and encompasses most concept explanation methods.

D Case Study

In this section, we present an illustrative case of the readability measures calculated via coherence-based measures and the LLM-based measure. We have the following observations.

First, extracted topics via highly activated contexts align well with and even exceed explanations generated by LLM (Fig. 7). As the number of samples inputted to LLM is restricted to a maximum context window and pricing limits (128,800 tokens and $0.03/1K tokens for GPT-4), explanations generated by LLM are only limited to the information presented. However, our coherency-based measures can search from a broader range of samples, looking for top-activating contexts to provide a more comprehensive explanation, as shown in Fig. 6.

Second, deep embedding-based measures are better at capturing semantic similarities. The first case illustrated in Fig. 6 (a) is ranked as the 1st among the 200 concepts evaluated by *IN-EmbCos* and 3rd by LLM, as it consistently activates on words related to geographical directions as suggested by LLM. However, *IN-UCI* only assigned a rank of 172. This is largely attributed to the fact that these terms may only occur once in a sample, showing one single direction, leading to low word co-occurrence counts.

Third, coherency-based measures can compensate for failure cases of LLM. For the 3rd case shown in Fig. 6, we can observe that it activates on expressions related to LaTeX. However, as LLM can only observe limited examples, it fails to include other attributes than mathematical symbols and markup, thus failing to simulate activations that align with the original activation. We approach this challenge by extracting topics from a larger range of samples.

Overall, these findings are consistent with the ones disclosed in Sec. 5.2.2, offering a more intuitive understanding of the measures' advantages and weaknesses.

E Sensitivity Analysis

In our sensitivity analysis of validity results, we expand beyond the examination of the 3rd layer of Pythia-70M, as depicted in Fig. 3. We include

Model	#Layer	#Param	#Dimesion
GPT-2 (small)	12	124M	768
Pythia-70M	6	70M	512

Table 6: Statistical model properties for subject models. #Layer, #Param, and #Dimension represent the number of layers, parameters, and dimensions respectively.

results from the 1st (Fig. 8 (a)) and 5th (Fig. 8 (b)) layers of Pythia-70M, as well as results from the 6th layer of GPT2-small (Fig. 8 (c)). Across these layers, reliability and validity results are consistent, with measures showing slightly better subset consistency in deeper layers. We speculate that as the layers deepen, the model discards irrelevant information and noise, leading to more stable and robust representations that are subject to less random error and exhibit higher consistency. Notably, the validity results on the 6th layer of GPT2-small align with our main findings (Fig. 8 (c)), fluctuating within a reasonable range, typically less than 0.1. These results underscore larger language models' superior ability and reliability compared to their counterparts, such as the 3rd layer of Pythia-70M.

F Implementation Details

In our implementation, we employ the Pile dataset, truncating input to 1024 tokens for efficient analysis. Both the Pile dataset and the backbones utilized are accessible for download from the Hugging Face Hub. To compute embedding-based readability measures, we leverage the backbone model's embedding matrix to extract token embeddings. All correlation metrics utilized in our analysis are calculated using the scipy package.

Following (Bills et al., 2023; Cunningham et al., 2023), we adopt the extraction of neuron activation as the output of the MLP layer in each layer, where each dimension corresponds to a neuron. Specifically, for a feed-forward layer $\text{FFN}(h_{in}) = \text{GeLU}(h_{in}W_1)W_2$, the MLP output/neurons are $\text{GeLU}(h_{in}W_1)$. Furthermore, the disentanglement-based baseline can utilize these extracted neurons as inputs to discover mono-semantic concepts, leveraging sparse autoencoders. We obtain the concept activation function of TCAV following (Kim et al., 2018). We treat LLM's harmful QA (Bhardwaj and Poria, 2023) as positive examples, and random texts as negative examples. Then, a linear classifier is trained on their hidden representations to classify harmful examples. The trained classifier's output function is regarded as the concept's

(a) Case 1 (b) Case 2 (c) Case 3

Figure 6: Topics extracted for calculating coherency-based measures. Spaces are replaced by '·' for visualization. These topics align well with LLM-generated explanations in Fig. 7 while providing fine-granular information.

Figure 7: A case study on LLM-based measure for readability measures. We present three cases with GPT4-generated explanation, original activation, and GPT4-simulated activation. GPT-4 performed well in the first two cases but worse in the third case.

activation function.

We employ a sparse autoencoder proposed by (Cunningham et al., 2023) to obtain concepts for the disentanglement-based baseline. The process involves running the model to be interpreted over the text while caching and saving activations at a specific layer, as narrated above. These activations then constitute a dataset used for training the autoencoders. The training is executed with the Adam optimizer, employing a learning rate of 1e-3, and processing 11 billion activation vectors for one epoch. The dictionary size is set at 8 times the hidden space's dimension. A single training run with this data volume is completed in approximately 13 hours on a single RTX 3090 GPU. To balance between sparsity and accuracy, we set the coefficient on the L1 loss to 0.5 for the 3rd layer of Pythia-70M.

It's important to note that our approach is in line with the original experimental setups outlined in (Bills et al., 2023; Cunningham et al., 2023; Kim et al., 2018). For a more detailed understanding of the implementation settings, interested readers are encouraged to refer to the original papers.

In calculating faithfulness, GRAD-Div is neglected as gradient operation is only applicable to one variable at a time, applying gradient operation to the whole output class is computationally expensive. To aggregate the effect on each token, they are weighted by their activations. Samples that exhibit high activation levels regarding a specific concept are deemed more relevant to the concept empirically and therefore receive higher weights. This weighting scheme ensures that the most representative samples contribute more significantly to the evaluation, enhancing the fidelity of the faithfulness measure in capturing the alignment between the model's behavior and the intended concept.

For LLM-based readability score (Bills et al., 2023), we follow OpenAI's pipeline as illustrated in (Bills et al., 2023). We show a detailed prompt for generating an explanation/semantic expression of a concept based on its activation in Fig. 9. We adopt this adjusted algorithm with `gpt-4-turbo-preview` as the simulator, due to new limitations in calculating logprobs on the input side. When extracting patterns that maximally activate a concept, we keep only the top 10 tokens with the largest activation or contribution to high-activation tokens.

G User study settings

In our user study, we recruited 3 human labelers to evaluate the readability of 200 concepts. The human labelers possess a high school level of English proficiency, allowing for easy comprehension of the concepts. These labelers were selected from within our academic institution to ensure a consistent educational background, which is pertinent to the readability aspect of our study. To maintain the quality of labeling, we implemented a compensation structure that rewards the labelers based on the number of concepts they evaluate. This approach was designed to incentivize thorough and careful consideration of each concept.

During the study, labelers were required to complete their assessments within a five-minute win-

A. Faithfulness

(a) 1st layer of Pythia-70M

	ABL-Loss	ABL-Div	ABL-TClass	ABL-PClass	GRAD-TClass	GRAD-PClass	IN-EmbDist	IN-EmbCos	OUT-EmbDist	OUT-EmbCos
ABL-Loss	0.51	0.52	0.54	0.44	0.19	0.19	0.19	0.28	0.22	0.03
ABL-Div	0.52	0.58	0.56	0.61	0.18	0.20	0.24	0.40	0.24	0.04
ABL-TClass	0.54	0.56	0.80	0.86	0.38	0.40	0.28	0.35	0.19	0.01
ABL-PClass	0.44	0.61	0.86	0.81	0.36	0.39	0.27	0.37	0.19	0.02
GRAD-TClass	0.19	0.18	0.38	0.36	0.91	0.89	0.10	0.12	0.08	-0.02
GRAD-PClass	0.19	0.20	0.40	0.39	0.89	0.91	0.10	0.12	0.08	-0.01
IN-EmbDist	0.19	0.24	0.28	0.27	0.10	0.10	0.98	0.62	0.27	0.12
IN-EmbCos	0.28	0.40	0.35	0.37	0.12	0.12	0.62	0.97	0.37	0.13
OUT-EmbDist	0.22	0.24	0.19	0.19	0.08	0.08	0.27	0.37	1.00	0.37
OUT-EmbCos	0.03	0.04	0.01	0.02	-0.02	-0.01	0.12	0.13	0.37	1.00

(b) 5th layer of Pythia-70M

	ABL-Loss	ABL-Div	ABL-TClass	ABL-PClass	GRAD-TClass	GRAD-PClass	IN-EmbDist	IN-EmbCos	OUT-EmbDist	OUT-EmbCos
ABL-Loss	0.99	0.64	0.31	0.29	0.09	0.06	0.09	0.17	0.18	0.10
ABL-Div	0.64	0.99	0.41	0.42	0.09	0.09	0.08	0.15	0.19	0.06
ABL-TClass	0.31	0.41	0.98	0.94	0.61	0.60	-0.09	0.01	0.27	0.12
ABL-PClass	0.29	0.42	0.94	0.98	0.61	0.62	-0.09	0.01	0.27	0.11
GRAD-TClass	0.09	0.09	0.61	0.61	0.99	0.91	-0.08	0.04	0.31	0.16
GRAD-PClass	0.06	0.09	0.60	0.62	0.91	0.99	-0.08	0.04	0.31	0.16
IN-EmbDist	0.09	0.08	-0.09	-0.09	-0.08	-0.08	0.90	0.64	0.03	0.06
IN-EmbCos	0.17	0.15	0.01	0.01	0.04	0.04	0.64	0.94	0.12	0.12
OUT-EmbDist	0.18	0.19	0.27	0.27	0.31	0.31	0.03	0.12	1.00	0.34
OUT-EmbCos	0.10	0.06	0.12	0.11	0.16	0.16	0.06	0.12	0.34	1.00

(c) 6th layer of GPT-2 small

	ABL-Loss	ABL-Div	ABL-TClass	ABL-PClass	GRAD-TClass	GRAD-PClass	IN-EmbDist	IN-EmbCos	OUT-EmbDist	OUT-EmbCos
ABL-Loss	0.96	0.36	0.27	0.23	0.07	0.05	0.14	0.19	0.19	0.09
ABL-Div	0.36	0.99	0.42	0.46	0.13	0.14	0.21	0.27	0.40	0.23
ABL-TClass	0.27	0.42	0.96	0.89	0.47	0.47	0.13	0.14	0.27	0.15
ABL-PClass	0.23	0.46	0.89	0.96	0.44	0.45	0.15	0.16	0.29	0.18
GRAD-TClass	0.07	0.13	0.47	0.44	0.99	0.92	0.03	0.02	0.08	0.01
GRAD-PClass	0.05	0.14	0.47	0.45	0.92	0.98	0.03	0.03	0.09	0.03
IN-EmbDist	0.14	0.21	0.13	0.15	0.03	0.03	0.91	0.47	0.17	0.16
IN-EmbCos	0.19	0.27	0.14	0.16	0.02	0.03	0.47	0.85	0.19	0.12
OUT-EmbDist	0.19	0.40	0.27	0.29	0.08	0.09	0.17	0.19	1.00	0.27
OUT-EmbCos	0.09	0.23	0.15	0.18	0.01	0.03	0.16	0.12	0.27	1.00

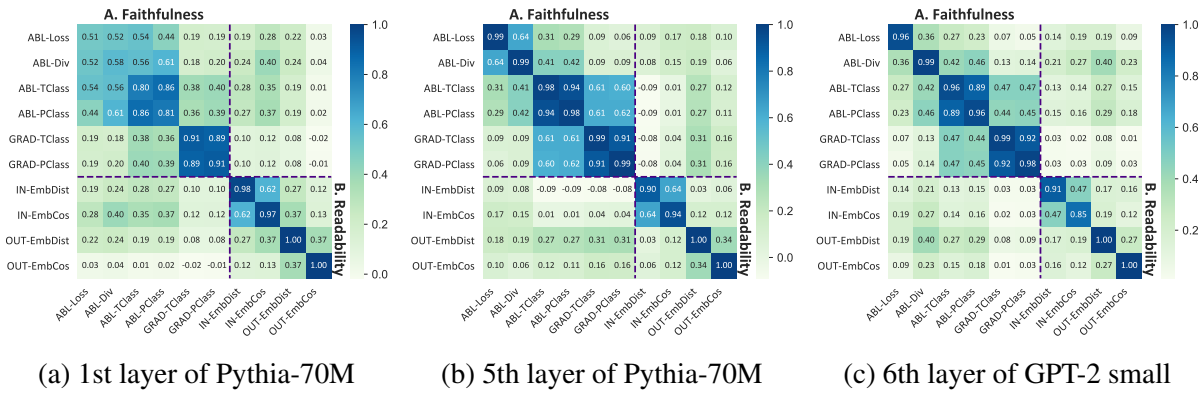*(B. Readability axis labeled along the right side of each subfigure.)*

Figure 8: The MTMM table of the evaluation measures: 1) subset consistency is shown on the diagonals; 2) construct validity is displayed on the off-diagonals.

dow for each concept. This time constraint was established to simulate a realistic scenario in which users make quick judgments about concept readability. Each of the three labelers was presented with the same set of 200 concepts to ensure consistency in the evaluation process.

Given input or output side tokens for a concept, each of our human labelers gives one readability score by simultaneously considering the three aspects, including semantic, grammatical or syntactic, and morphological information. More specifically, a concept is considered highly readable, if it is related to a specific topic such as computer systems (semantically interesting), is associated with a specific grammar or syntax (grammatically or syntactically interesting), or consists of tokens that share a similar structure or a form such as all being usable as suffixes for a certain token (morphologically interesting).

We demonstrate guidelines that were provided to the labelers. These guidelines were crafted to assist the labelers in their task and to standardize the evaluation criteria across all participants. The guidelines are as follows:

Welcome to the user study on evaluating the readability of concepts extracted from concept-based explanations. Your valuable insights will contribute to advancing our understanding of these explanations and improving their interpretability. Below are the instructions for scoring each concept:

Task Overview. *You will be provided with a list of concepts, each comprising three parts:*

- *Activation of this concept in 10 sentences, with each sentence containing 64 tokens.*

- *The 20 tokens that have the greatest impact on its activation value.*

- *The model's output of the 20 tokens with the highest logits after replacing hidden states with the direction of the concept.*

For each concept, please provide two scores within the range of $[1, 2, 3, 4, 5]$, *representing their perceived readability of the relevant information on the input and output sides.*

Evaluation Criteria. *Please consider the following aspects when scoring each concept:*

- *Semantic Information: Consider whether the concept is related to a specific topic, such as containing terms related to computer systems.*

- *Grammatical or syntactic information: Assess whether the concept is associated with specific grammar or syntax, such as being frequently activated with various copulas.*

- *Morphological Information: Consider whether the given tokens share a similar structure or form, such as all being usable as suffixes for a certain token.*

Scoring Procedure. *Please provide a score for the input side, reflecting the readability of tokens related to the concept in the input. Additionally, assign a score for the output side, indicating the readability of tokens related to the concept in the output. Your engagement in this scoring procedure will significantly contribute to the comprehensiveness of our study. Thank you for your participation!*

624

Input

```
<SYSTEM>: We're studying neurons in a neural network. Each neuron looks for some particular
thing in a short document. Look at the parts of the document the neuron activates for
and summarize in a single sentence what the neuron is looking for. Don't list
examples of words.\n\nThe activation format is token<tab>activation. Activation
values range from 0 to 10. A neuron finding what it's looking for is represented by a
non-zero activation value. The higher the activation value, the stronger the match.

<USER>: Neuron 1
Activations:

<start>
...
warm 2
ly 0
affecting 6
and 0
so 0
...
<end>

Same activations, but with all zeros filtered out:

<start>
...
warm 2
affecting 6
...
<end>

Explanation of neuron 1 behavior: the main thing the neuron does is to find:

<ASSISTANT>: present tense verbs ending in 'ing'

...more few-shot examples...

<USER>: Neuron 4:
Activations:

<start>
...
information 2
systems 0
infrastructures 0
including 0
the 0
Internet 4
, 0
telecommunications 4
networks 6
...
<end>

Same activations, but with all zeros filtered out:

<start>
...
information 2
Internet 4
telecommunications 4
networks 6
...
<end>

Explanation of neuron 4 behavior: the main thing the neuron does is to find:

<ASSISTANT>:
```

Output

```
Terminologies related to computer networks
```

Figure 9: Prompt and example input and output for generating a semantic expression for a given concept (Bills et al., 2023). '\t' is used as the separator between a token and an activation value.