

Evaluating BERT on Question Exhaustivity Inference

Stanford CS224N Custom Project

Dhara Yu

Department of Computer Science
Stanford University
dharakyu@stanford.edu

Abstract

Pragmatic listeners possess the ability to infer the intended exhaustivity of a question, even when the question is semantically underspecified. For example, in response to the question *Where can I get coffee around here?*, listeners tend to infer the question-asker is requesting a non-exhaustive list of coffee places. In this work, we explore the extent to which BERT can learn such exhaustivity judgements. We first show that BERT, finetuned on a small dataset of questions and human judgements of exhaustivity modeled as a probability distribution over 4 categories, can predict these judgements with high accuracy ($r = 0.65$, $r = 0.59$, $r = 0.76$ and $r = 0.7$ across each category). We also provide evidence that the model learns associations between the linguistic features identified as modulating exhaustivity judgement, and the magnitude of the judgement.

1 Key Information to include

- Mentor: Elissa Yi, Judith Degen (external)
- External Collaborators (if you have any):
- Sharing project: No

2 Introduction

Humans have the ability to infer information beyond the literal meaning of an utterance, by implicitly integrating contextual information and linguistic cues. This ability, known as pragmatic reasoning, can be seen in question response tasks, where listeners must recover the speaker's intended meaning, given the discourse context and the question asked. As an illustration of this phenomenon, consider the following questions:

1. *Where can I get coffee?*
2. *Who are the people in charge here?*

Both utterances are semantically underspecified for exhaustivity at the surface-level; that is, one cannot recover from signal alone whether a speaker responds by providing *all* (an exhaustive response) or *some* (a non-exhaustive response) of the items that fit the criteria. Despite this underspecification, there still appear to be inherent differences in whether a question is interpreted exhaustively or non-exhaustively. A pragmatic listener who hears (1) has a natural tendency to answer the first by naming one or a few nearby coffee shops because they assume the questioner's goal is non-exhaustive, and answer the second by providing an exhaustive list of people in charge. However, in the first situation, if the answerer becomes aware that the questioner is a professional coffee critic, they might be inclined to provide an exhaustive list of coffee shops instead, because it is compatible with the critic's goals to review *all* the local cafes. This example illustrates the role that contextual information plays in modulating judgement [1] [2].

At the same time, some have noted that linguistic differences could be at work in determining question interpretation. For example, the presence of a modal auxiliary as in (1) is said to specially license a non-exhaustive interpretation, while some researchers cite differences due to the *wh*-word.

The theoretical linguistics literature assumes that there is little variability in question interpretation, and that questions are interpreted exhaustively by default with the exception of specific linguistic features. However, a recent corpus study [3] has shown that there is more variability than previously assumed, and in fact that questions are not interpreted exhaustively by default.

In this paper, we explore the extent to which BERT, a pretrained transformer language model that has achieved state-of-the-art performance on a host of NLP tasks but lacks explicit pragmatic reasoning ability, can learn to predict exhaustivity judgements of *wh*-questions. To do so, we finetune on the small dataset of naturalistic *wh*-questions and accompanying human judgements of exhaustivity from [3], and assess the strength of alignment between the model predictions and human judgement. We find that the best finetuned BERT model predicts exhaustivity on a held-out test set with high accuracy, yielding correlation coefficients of 0.65, 0.59, 0.76 and 0.7 across the categories of interest. We also find preliminary evidence that the model learns associations between the predicted values and the linguistic features identified as modulating exhaustivity judgements, as reported in [3, 1].

Although question exhaustivity is not as widely studied as other forms of pragmatic inference, it nonetheless serves an important benchmark for pragmatic reasoning ability due to its ubiquity in naturalistic speech. Furthermore, we see question exhaustivity as a topic of particular interest for applications such as dialogue agents. For example, having the ability to reason about the desired number of suggestions by using both linguistic and contextual information to reconstruct user goals, could significantly increase the conversational fluency of an agent.

3 Related Work

3.1 Ability of language models to learn pragmatic inference tasks

In light of the success of neural network language models on various natural language tasks, some recent works have explored the ability of such models to learn different types of pragmatic inference by training on a small dataset of pragmatic task inputs and outputs, and evaluating alignment of the model’s prediction with human judgement.

Within the broader category of pragmatic tasks, there appears to be a difference in performance between implicature and presupposition tasks. Schuster et al. [4] utilize a crowdsourced dataset of scalar implicature judgements [5] which is generated by asking humans to rate on a Likert scale the extent to which they agree that a given sentence with a *some* descriptive entails *not all*. They find that an LSTM model trained on this dataset can predict judgements of scalar inference from *some* to *not all* with high accuracy, and that the model learned associations between judgements and the linguistic features previously reported as modulating scalar implicature.

However, White et al. [6] and Jiang and de Marneffe [7] report systemic errors in trained models’ ability to perform pragmatic presupposition tasks. They train LSTM models and BERT, respectively, on crowdsourced datasets of presupposition tasks, where humans are asked to judge how strongly an embedded clause suggests that the complement clause is true. Both find that their models fail to integrate appropriate linguistic cues in making judgements of factuality.

Even within the category of scalar implicature, language models appear to be more equipped to learn certain types of implicatures better than others. Li et al. [8] employ the architecture from [4] and observe that the same LSTM models, trained on a different crowdsourced dataset of scalar implicature judgements from *or* to *but not both*, are unable to accurately predict judgements of this type of scalar implicatures.

Other papers have attempted to take a more birds-eye perspective of the inferential abilities of language models. Jeretic et al. [9] find that BERT, finetuned on an assorted corpus of natural language inference tasks, can perform certain pragmatic tasks such as treating *some* in scalar implicatures as an entailment, as well as recognize a presupposition as an entailment in the presence of certain trigger words. Ettinger [10] evaluates BERT with no finetuning on several psycholinguistic diagnostic tests and observes that for a sentence completion task requiring pragmatic reasoning ability, BERT

Speaker #2: pretty good.
Speaker #1: i do like to ski.
Speaker #2: pretty, pretty down there. huh?
Speaker #1: yeah, i , i said i do like to ski.
Speaker #2: *so, where have you skied?*

Based on the sentence in red, how likely do you think it is that the spaker wanted to know about each of the following?

What is every place...?		0
What is a place...?		0
What is the place...?		0
Something else		0

Continue

Figure 1: Example prompt shown to crowdworkers to generate *wh*-question dataset. Participants asked to distribute 100 points representing their agreement with a paraphrase.

provides a sensible answer about half the time, so overall model performance falls short of human ability.

This paper aims to further elucidate the pragmatic abilities of language models by probing the performance of BERT on a novel pragmatic inference task.

3.2 Empirical observations of *wh*-question exhaustivity judgement

Moyer and Degen [3] explore the factors that modulate human judgement of *wh*-question exhaustivity by performing a naturalistic corpus study. To compile the corpus, participants were presented with a naturalistic *wh*-question and 10 lines of preceding dialogue. Participants then rated the likely intended meaning - whether the question suggested mentioning *every* item, *a* single item, *the* item, or none of the above - using a continuous slider for each paraphrase (Figure 1). The total slider points across all paraphrases summed to 100 to ensure that the ratings represented a complete probability distribution. Moyer and Degen find a strong bias for *the* paraphrases, and that judgements are modulated by linguistic form, specifically modality and choice of *wh*-word.

We use this novel task formulation and accompanying dataset as the ground truth to assess our model’s ability to learn *wh*-question exhaustivity.

4 Approach

We use the Huggingface pretrained implementation of BERT [11], the state-of-the-art model for natural language inference tasks, to finetune on the dataset of *wh*-question exhaustivity. We add a classification head comprised of additional layers to output a probability over 4 classes. Specifically, we implement a linear projection layer from the hidden outputs, followed by a ReLU activation layer, followed by a dropout layer, and finally a linear classification layer. We then use the softmax function to output a probability distribution that sums to 1.

To train models and run the experiments we implement a new codebase, using code samples from Huggingface tutorials and CS224N Assignment #5. We also write scripts to preprocess the raw empirical data from [3]; specifically, we clean annotation data from the questions and preceding dialogue, and split the dataset into training, validation and test. Finally, we write code to evaluate the held-out test set on our models, and analyze model performance. All code and datasets are available here: <https://github.com/dharakyu/wh-questions-lm>.

Since this is, to our knowledge, the first attempt to study pragmatic inference of question exhaustivity in language models, we elect to use a simple and relatively unsophisticated baseline to make our results more immediately interpretable. The baseline is as follows: we compute the mean of each judgement rating in the dataset, and compute the Earth Mover’s Distance (described in more detail in

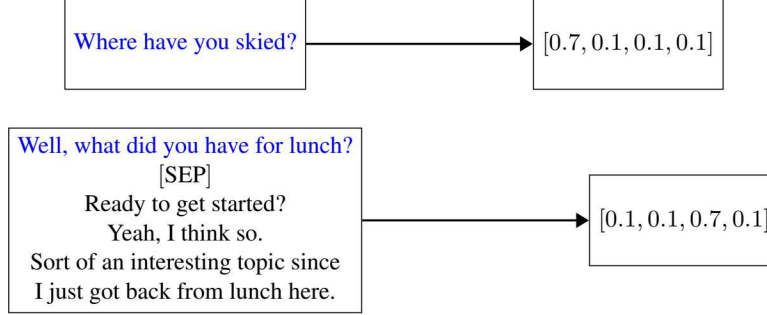


Figure 2: Examples of input and output data for the prediction task. The output vector represents the participants certainty that the questioner wants to know *every* place the listener has skied, *a* place, *the* place, or something else. The first example is with the question only, and the second example shows the question concatenated with the preceding context.

the following section) between the mean rating and each true label. We then take the mean of those distances as the baseline for the test set.

5 Experiments

5.1 Data

We use the dataset released by [3] (See Figure 1 for an example of the prompt used to collect data from participants). Questions and preceding dialogues in the dataset are raw speech transcriptions from another corpus, so we perform preprocessing to remove unnecessary characters and annotations. As multiple participants responded to each question, we perform additional preprocessing to compute the average distribution for each question. This gives a total of 995 unique questions and paraphrase distributions. Of that, we randomly sample 697 data points to be used for training and validation, and hold the remaining 298 to be used as the test set.

Our model task is: given a question, predict the certainty of classification as a paraphrase, over 4 classes. For example, given the question *Where have you skied?*, the goal is to predict the certainty that 1) the speaker who asked the question is expecting *every* place the person has skied, 2) *a* place, 3) *the* place, and 4) something else. We also try a modified version of this task, where the model is given the question and the preceding context, demarcated by a BERT [SEP] token (Figure 2).

5.2 Evaluation method

To evaluate how closely the judgements output by the model adhere to the empirical data, we use the Earth Mover’s Distance metric [12]. Intuitively, this metric measures the distance between two probability distributions by computing the minimal amount of work to transform one distribution into the other. We use the SciPy implementation to compute EMD between each prediction and the label, and then compute the mean of distances to get an aggregated distance metric.

5.3 Experimental details

We finetune BERT on the training dataset with the questions only, and with the questions and preceding contexts appended. We use the Adam optimizer with learning rate 5e-06 to minimize Kullback–Leibler divergence between the predicted probability distribution over paraphrases and the labels. For our dropout layer in the linear classifier head, we use dropout rate of 0.3. We train for 60 epochs on a single Tesla K80 GPU, with batch size of 32.

5.4 Results

Figure 3 provides a summary of model performance. We evaluate performance on the test set with the question only, and with the question and preceding context. Generally, we find that low dropout rate

Model	Question only	Question + preceding context
Baseline	0.107	0.107
BERT, question only	0.075	0.118
BERT, question + preceding context	0.104	0.087

Figure 3: Model performance on held-out test set. Metric used is Earth Mover’s Distance.

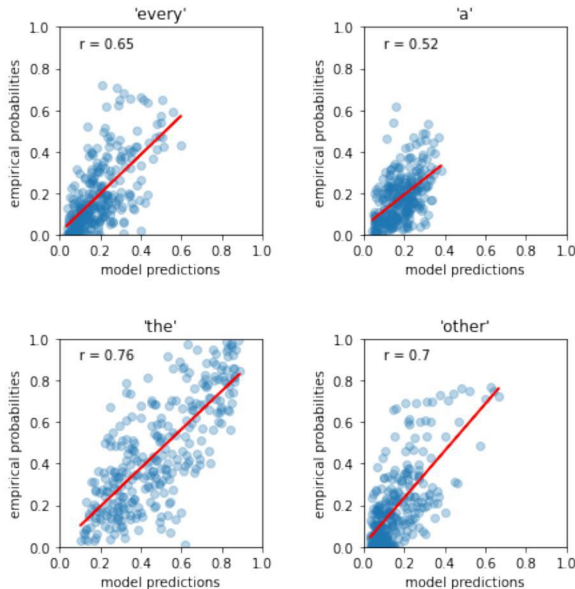


Figure 4: Correlation between predictions of the BERT model trained on the question only dataset, and the empirical probabilities, across all 4 categories of paraphrases.

(as in high probability of dropping an input) and small learning rate most significantly improve model performance. Among the 4 combinations, we observe that the model trained only on the question, and evaluated on the question-only test set, performs the best. This appears to replicate the results from Schuster et al, as they found that including the preceding conversational context during training did not improve or lowered the accuracy of the prediction [4]. See section 6.3 for a more in-depth analysis of context.

Unsurprisingly, we observe that model performance degrades when the input data from the test set is not of the same format as the input data from the training set, although it appears that the model trained on the question and the preceding context is somewhat more robust to the change in inputs.

While EMD is a useful summative metric to help us understand relative performance, it by itself does not yield much insight into the extent that the model predictions correlate with the the empirical data. Accordingly, we plot the predicted probability for our best-performing across each paraphrase category (*every*, *a*, *the*, and *other*) against the empirical probability in the dataset, and compute the Pearson correlation coefficient for each (Figure 4). We observe that the model predicts the judgements of *every* and *a* with moderate accuracy ($r = 0.65$ and $r = 0.59$), and predicts judgements of *the* and *other* paraphrases with high accuracy ($r = 0.76$ and $r = 0.7$). The primary systemic deviation appears to be that the model tends to underpredict high empirical probabilities.

Overall, these results suggest that the model learned to predict exhaustivity judgement fairly well. With that said, this does not establish that the model learned the relationships between linguistic features identified in [3] as modulating exhaustivity, and the prediction. In the next section, we explore this further.

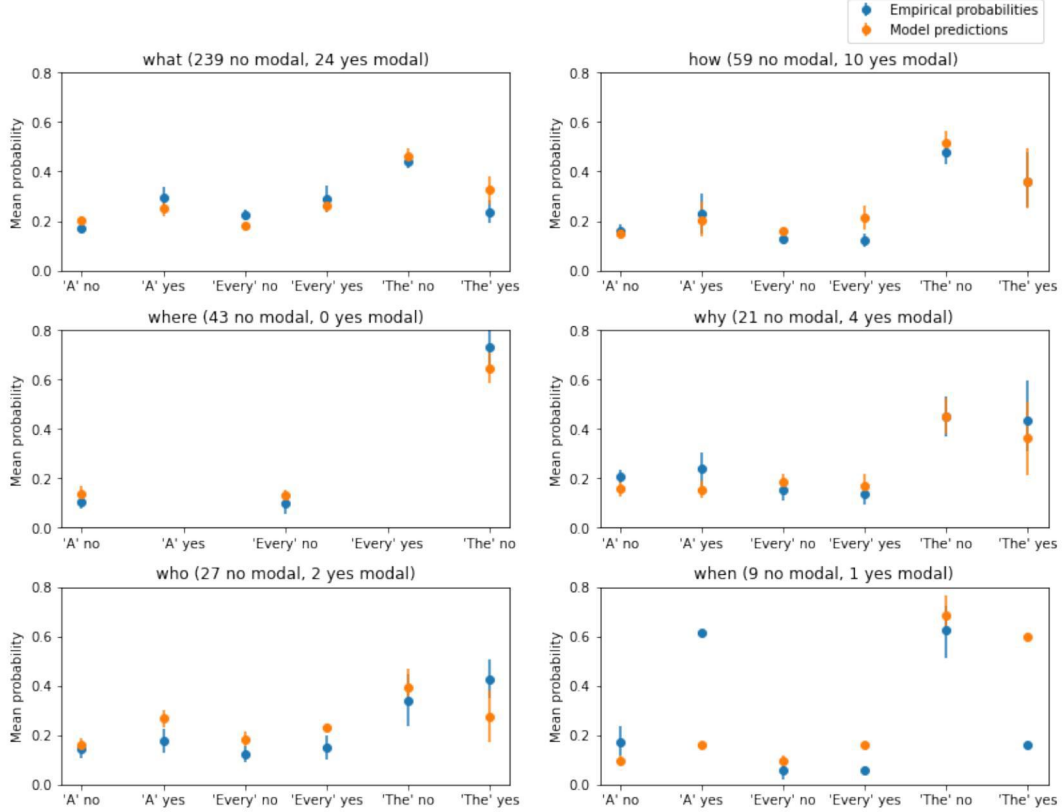


Figure 5: Mean empirical probabilities and model predictions, across *wh*-word. *no* indicates no modal present and *yes* indicates a modal present. Number of examples for each category of *wh*-word, broken down by modality, appears at the top of each plot. Error bars indicate 95% bootstrapped confidence intervals.

6 Analysis

6.1 Measuring association between linguistic features and prediction

Moyer and Degen [3] identify *wh*-word and modality as linguistic factors, which have been discussed in the theoretical literature as affecting question interpretation. Consider again examples (1) and (2), which differ in *wh*-word and presence/absence of the modal auxiliary *can*. The ‘who’ question is thought to be unambiguous, while the ‘where’ question could be interpreted in an exhaustive or non-exhaustive way.

Minimal pair analysis, in which artificial sentences are constructed to differ along a linguistic feature and the resulting model results are compared, is one method that has been effectively used to show model sensitivity to different inputs [4]. However, minimal pair analysis for our dataset is challenging because in many cases, swapping in a different *wh*-word results in ungrammatical sentences that do not make semantic sense (i.e. *Where can I get coffee?* → *What can I get coffee?*). Accordingly, we try to approximate minimal pair analysis with the following approach: we visualize how the mean magnitude of prediction varies with *wh*-word and modality, for each category of *every*, *a* and *the* paraphrases, and compare it to the empirical data. Since several *wh*-word and modality combinations occur very few times, we augment the evaluation dataset by combining the validation set with the test set, which is acceptable as the model has not been trained on either. We provide a breakdown of the number of occurrences of the given combination (Figure 5).

First, we observe that the model replicates the empirical observations of the strong bias for *the* paraphrases observed in [3]. This is not surprising, as the *the* bias would have been learned during training. More strikingly, we see that model’s mean prediction for a given *wh*-word and modal

Sentence	Model prediction	Empirical probabilities
what do they,	[0.10, 0.12, 0.40, 0.38]	[0.05, 0.05, 0.17, 0.73]
why don't you tell me about yours	[0.21, 0.26, 0.31, 0.22]	[0.09, 0.20, 0.10, 0.61]
why don't you start	[0.15, 0.22, 0.29, 0.34]	[0.02, 0.10, 0.17, 0.70]
why don't you go first	[0.21, 0.21, 0.27, 0.31]	[0.11, 0.10, 0.09, 0.70]
and then, uh, who knows how they process it from there and what they turn it into	[0.21, 0.26, 0.24, 0.28]	[0.08, 0.22, 0.09, 0.60]

Figure 6: Example sentences that yielded a large distance between the model prediction and the empirical data. The first value in the probabilities vector corresponds to *every*, the second to *a*, the third to *the*, and the fourth to *other*.

combination closely tracks the empirical mean. It should be emphasized that the data samples for certain combinations are very limited. For example, there is only one instance of a modal *when*-question. However, for the combinations with more numerous data points, the correlation appears to be fairly strong. For example, we see that the model learns that for *what* questions, the presence of a modal correlates with a higher probability of an *every* or *a* paraphrase and a lower probability of a *the* paraphrase.

Overall, these results provide evidence that the model learned associations between the linguistic features of interest and the prediction of exhaustivity.

6.2 Qualitative analysis of model outputs

We qualitatively examined the questions that resulted in the worst model performance. We find that many of the questions for which the model failed to predict exhaustivity had high empirical *other* ratings, indicating that they were rhetorical in nature, such as *why don't you start*, or otherwise could not be classified as exhaustive or non-exhaustive in meaning (Figure 6).

We also find several examples where the model correctly predicted the relative ordering, but the magnitudes were off, resulting in a relatively large gap between prediction and label. For example, for the question *who's your favorite actress or actor*, the model correctly predicted that *the* had the highest probability, followed by *a*, *every* and *other*, but it underestimated the magnitude of the *the* paraphrase, which significantly affected the EMD.

6.3 Further discussion of the role of preceding contextual dialogue

As noted earlier, we observed that the model performed worse on the EMD metric when it was trained with the preceding context in addition to the question. For the *every*, *a*, *the* and *other* paraphrases, the correlation coefficients were 0.55, 0.39, 0.6 and 0.5, respectively, which is a significant decrease from the model trained without context. That can be explained by two possibilities: 1) context is not needed to make exhaustivity judgements, and thus incorporating it into the model introduces distracting noise, or 2) the model failed to incorporate context. While 1) seems somewhat unlikely, at this point we cannot rule out either possibility. Because the dataset we use does not provide empirical probabilities when participants only had access to the question itself, we do not know how context changes the interpretation of exhaustivity on a systematic basis, if at all.

7 Conclusion

In this work, we show that BERT, finetuned on a small dataset of questions from naturalistic speech, learns to perform the pragmatic task of predicting question exhaustivity with high accuracy. Our analysis of model behavior provides preliminary evidence that BERT learns the associations between the linguistic features that modulate exhaustivity judgement, and the judgement itself. Qualitative analysis reveals that a substantive part of model inaccuracy can be attributed to highly ambiguous questions that defied classification for human participants.

Because of the condensed time frame of this project, we were not able to explore additional analyses to probe the extent to which the model actually learns linguistic features, beyond just the statistical

association. Jiang and de Marneffe provide a useful method for evaluating the encoding of linguistic features through freezing BERT layers [7], which we hope to emulate in future work. Furthermore, the dataset and the task formulation we worked with in this project is brand-new and have not been used in any other works, so there is more work to be done to understand any idiosyncracies that could explain our results. Lastly, we hope to shed more light on the question of how much context informs exhaustivity judgements by performing a replicate study with human participants, but without the 10-preceding lines of dialogue, to provide a more direct point of comparison with our model.

References

- [1] Morgan Moyer and Kristen Syrett. (Non-)exhaustivity in embedded questions: Contextual, lexical and structural factors. In M. Teresa Espinal, E. Castroviejo, M. Leonetti, L. McNally, and C. Real-Puigdollers, editors, *Proceedings of Sinn und Bedeutung 23*, pages 207–224, 2019.
- [2] Morgan Moyer. *The question of questions: resolving (non-) exhaustivity in wh-questions*. Doctoral dissertation, Linguistics Department, Rutgers University, 2020.
- [3] Morgan Moyer and Judith Degen. Who thinks wh-questions are exhaustive?, 2021. [10.31234/osf.io/ky5mn](https://doi.org/10.31234/osf.io/ky5mn).
- [4] Sebastian Schuster, Yuxing Chen, and Judith Degen. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
- [5] Judith Degen. Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8(11):1–55, May 2015.
- [6] Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
- [7] Nanjiang Jiang and Marie-Catherine de Marneffe. Evaluating BERT for natural language inference: A case study on the CommitmentBank. 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)*, 2019.
- [8] Elissa Li, Sebastian Schuster, and Judith Degen. Predicting scalar inferences from or to not both using neural sentence encoders. In *Proceedings of the Society for Computation in Linguistics*, 2021.
- [9] Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models IMPPRESSive? Learning IMPLICature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
- [10] Allyson Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8, 2020.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, 2019.
- [12] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover’s distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, November 2000.