Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes

Anonymous ACL submission

Abstract

Multi-hop Question Answering (QA) is a challenging task since it requires an accurate aggregation of information from multiple context paragraphs and a thorough understanding of the underlying reasoning chains. Recent work in multi-hop QA has shown that performance can be boosted by first decomposing the questions into simpler, single-hop questions. In this paper, we explore one additional utility of the multi-hop decomposition from the perspective of explainable NLP: to create explanation by *probing* a neural QA model with them. We hypothesize that in doing so, users will be better able to construct a mental model of when the underlying QA system will give the correct answer. Through human participant studies, we verify that exposing the decomposition probes and answers to the probes to users can increase their ability to predict system performance on a question instance basis. We show that decomposition is an effective form of probing QA systems as well as a promising approach to explanation generation. In-depth analyses show the need for improvements in decomposition systems.

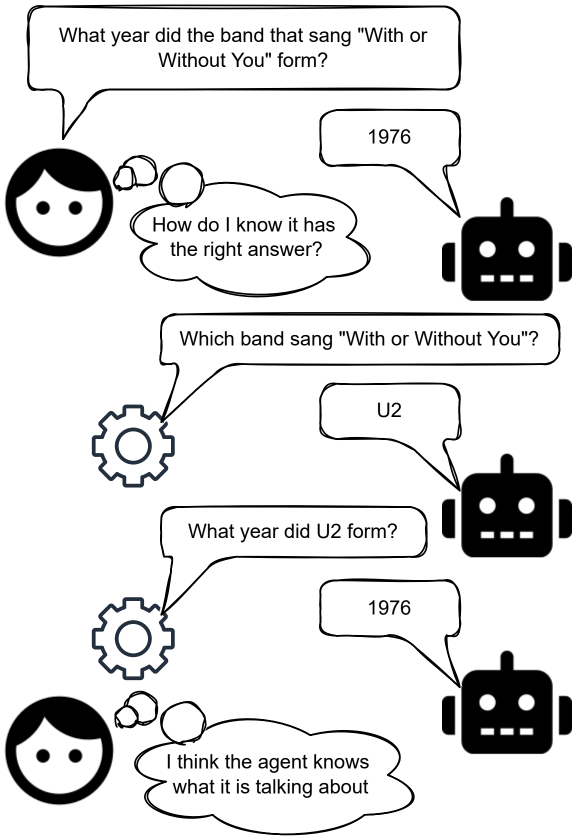

Figure 1: An overview of our method. The user wonders if they are able to trust the answer. Sub-questions are generated by a decomposer agent (gear) to probe the question-answering agent.

1 Introduction

As natural language understanding tasks have become increasingly complex, the field of explainable natural language processing (exNLP) aims to help users understand the performance of NLP systems. Multi-hop question answering is one such task in which questions seemingly require multiple reasoning steps to answer. To accurately answer a multi-hop question, one must start by decomposing the given multi-hop question into simpler sub-questions, then try to answer them respectively, and finally aggregate together the information obtained from all the sub-questions. For instance, consider the multi-hop question “What year did the band that sang ‘With Or Without You’ form?”. To answer the question, one must first figure out the band that sang that song from and then find the year in which that band formed from another context paragraph. A typical approach to multi-hop QA systems is to automatically decompose the question into sub-questions, answer those questions, and then synthesize the answers to the sub-questions to answer the original question (Min et al., 2019; Perez et al., 2020; Khot et al., 2021).

From the perspective of explainable NLP, we explore the utility of multi-hop decompositions to create explanations. One role of explanations is to help users construct a *mental model* of the underlying system (Chandrasekaran et al., 2017; Chakraborti et al., 2019; Jacovi et al., 2022). In doing so, users will be better equipped to know when the system answers can be trusted. This is especially important for large, general-purpose QA systems that can answer a wide range of questions but might have greater competencies when answering questions about some topics versus others . We hypothesize that question decompositions used to *probe* a neural QA model can improve users' abilities to predict whether the QA system will answer the original question correctly or not.

Khot (2021) observed that improved decompositional reasoning chains for multi-hop QA correlate with increased user perceptions of trust, understandability, and preference. While perceptions of trust are important, it is also important that the trust is *appropriately calibrated* (Perkins et al., 2021). That is, the user should trust the system when it is worthy of that trust. For general-purpose question-answering systems built upon large-scale language models the ability to accurately answer a question is likely to be variable based on the specific question asked.

How does a user know when to trust a QA system's answer to a particular question? If just presented with an answer, one has no cues from which to make an assessment. End-to-end QA systems that generate answers and explanations are trained to justify the answer as opposed to provide evidence of the system's competencies on a topic.

We introduce *probing* as an explanation strategy that helps a user determine whether to trust an answer. Probing is a process whereby a model is provided similar inputs to determine if its performance is stable when handling related inputs. In this work, we show that exposing the decomposition probes and answers to the probes to users can increase their ability to predict whether the system will answer the original question correctly. This indicates that users—without knowing the particulars of the underlying QA system—are receiving *actionable* cues from which to model the behavior of the system. Instead of asking for subjective perceptions of the overall system, we objectively measure the effect of the probes on instance-level interactions.

To the best of our knowledge, this paper is the first to show that probing can have a measurable impact on users in multi-hop QA. These results are also complementary to Tang et al. (2021) who

Context: Learning, Inc. is an educational software and hardware company co-founded in 1999 by Texas businessman Neil Bush and a year later Ken Leonard. He is the fourth of six children of former President George H. W. Bush and Barbara Bush (née Pierce).
Question: Who is the mother of the Texas business man that co-founded Ignite! Learning, Inc?
Answer: Barbara Bush

Sub-question 1: Who is the Texas business man who co founded Ignite Learning, Inc?
Answer: Neil Bush
Sub-question 2: Who is Neil Bush's mother?
Answer: Barbara Bush

Table 1: Example from the validation set of HotpotQA (Yang et al., 2018), as well as the associated silver question decompositions from Khot et al. (2021).

use decompositions to assess whether multi-hop QA systems successfully go through multiple hops when answering questions. In summary, our main findings are:

1. Decomposition is an effective form for probing neural QA models.
2. Explanation created by probing the neural QA model with question decompositions can help human construct a mental model on which they can rely to predict the model behavior.
3. Quality of decompositions matters—from the explainability perspective, existing question decomposers still have a long way to go.

A summary of our method is given in Figure 1.

2 Preliminaries

2.1 Dataset

We use a popular English question-answering / reading comprehension task designed to test multi-hop reasoning: HotPotQA (Yang et al., 2018). Examples are given in Table 1. The HotPotQA task involves answering questions by finding information over multiple Wikipedia articles.[1]

2.2 Question Decompositions

As a source of high-quality question decompositions and answers, we use the sub-questions and answers provided for a subset of the HotpotQA validation set by Khot et al. (2021).[2] These sub-questions are generated using distant supervision in the form of task-specific hints to a BART-Large

[1] We make simplifying assumptions for this task, detailed in §2.3.

[2] https://github.com/allenai/modularqa

(Lewis et al., 2020) model trained to generate questions in the SQuAD 2.0 dataset (Rajpurkar et al., 2018). The answers are generated by a RoBERTa-Large (Liu et al., 2019) model trained on SQuAD 2.0. These silver sub-question-and-answer pairs are relatively high-quality, in that the authors are able to use them to train a next-question generator that achieves high task performance on HotPotQA as part of a larger modular system.

All instances in the validation set have at least two sub-questions; certain questions have a third math operation sub-question that we abandon as this format only suits models with a numerical reasoning module (i.e., not conducive to being asked as a probe to the language model). The authors sampled 5 chains of sub-questions for each instance and filtered out noisy ones; we select the first from the remaining chains for our probes, and find that overall, the sub-questions and answers are of high quality and do not vary much across samples. This results in 676 instances for HotPotQA that have a silver question decomposition. Examples of question decompositions are given in Table 1.

Our choice of tasks is motivated by two factors: the existence of high-quality question decompositions and answers, and the task labels are not limited to predefined categories (such as *yes/no*), which limits the outputs that a fine-tuned generative model can produce when probed with sub-questions (i.e., if the dataset only contains yes/no questions, a model trained on it is unlikely to be able to answer sub-questions with anything other than yes/no). Future work can focus on extending fine-tuning protocols to apply the sub-question probing method to datasets with categorical labels.

2.3 Models

We fine-tune two popular pretrained models to perform the multi-hop QA tasks: T5-Base (220M parameters; Raffel et al., 2020) and BART-Base (140M parameters; Lewis et al., 2020). Both models are built on text-to-text encoder-decoder Transformer (Vaswani et al., 2017) architectures pretrained with denoising objectives. Both models treat question-answering tasks as generation tasks, making them well-suited for probing since they can thus also answer sub-questions in free-form natural language (rather than predicting from a fixed set of classification labels). We fine-tune the models using standard cross-entropy loss to generate the answer given the question and context. While one subtask for HotPotQA is to *select* the relevant context, i.e., the supporting paragraph from which to extract an answer, we focus on general architectures that are not designed for retrieval. Therefore, we provide the gold context paragraph as input. More details, including input-output formatting, are given in Appendix A.

	Metric		Metric (On Subset)		
Model	**EM**	**F1**	**EM**	**Manual**	**F1**
T5	66.73	79.97	70.27	91.27	85.41
BART	62.21	76.18	65.98	88.31	82.12

Table 2: Task performance of pretrained models on the validation set and a subset of it (see §2.4). "Manual" indicicates our manual annotation for answer correctness, which is more accurate than EM. A comparable model on HotPotQA (Tu et al., 2020) achieves 61.32 EM and 74.81 F1 on the full validation set.

The HotPotQA leaderboard relies on two metrics for determining answer correctness, originally from SQUAD (Rajpurkar et al., 2016): exact match (EM), whether the predictions and ground-truth answers match exactly, and F1 score, the (macro-averaged) token-level overlap between a prediction and ground truth answer (treating both as a bag-of-tokens). Using gold context paragraphs, our models achieve comparable performance to standard baselines on the answering task, reported in Table 2. T5 outperforms BART on both metrics. Our goal is *not* to build the best model but to establish a model with sufficient performance on questions and sub-questions to test our hypotheses about the effect of question decompositions as explanations.

Crucially, we never fine-tune on sub-questions. This allows our probing method to represent what a model that is *only trained for the task* knows about the task, without introducing any new information that may shift the predictions of the model in favor of the explanation fine-tuning corpus (such as is done in prior work; Roberts et al., 2020).

2.4 Probing with Sub-Questions

Given the fine-tuned models, we probe on a subset of the validation instances for which we have silver sub-questions—676 instances for HotPotQA. This is done at inference-time following the same format as the main task, i.e., by feeding each sub-question for an instance with the instance's gold context as input to the trained model. For each instance in the dataset, this process results in a tuple of the form: main question (Q), gold context paragraph (C), the model's predicted answer to the main question (A), two silver sub-questions (SUB-Q_1 and SUB-Q_2), and the model's predicted answers to the sub-questions (SUB-A_1 and SUB-A_2).

To avoid bias introduced by requiring an exact token match or determining an F1 cutoff for correctness of a model answer, we manually annotate the instances (both main and sub-questions) for correctness. This leads to a slight increase in accuracy due to instances where EM=0 but we determine the predicted answer to be correct (e.g., the correct answer is "Nashville", and the model predicted "Nashville, Tennessee"). For an example of how accuracy numbers change as the result of manual annotation, see the 4^{th} and 5^{th} columns of Table 2.

2.5 Simulatability

To understand how *faithful* explanations are to the underlying model as reflected by the mental model humans can develop of a machine learning system, the explainable AI community has long turned to **simulatability** experiments (Kim et al., 2016; Doshi-Velez and Kim, 2017; Ribeiro et al., 2018; Nguyen, 2018; Chandrasekaran et al., 2018; Hase and Bansal, 2020, *inter alia*). Doshi-Velez and Kim (2017) define "forward simulation/prediction" as the task by which "humans are presented with an explanation and an input, and must correctly simulate the model's output". They class this as a form of *human-grounded evaluation*, which has strengths over automatic evaluation methods because it investigates the understanding of real human users, and thus tests the utility of explanations in settings closer to true applications. Simulatability, to date, is one of the only human-grounded evaluation methods that tests the *interpretability* of explanation methods rather than *human preferences*, and is the most widely used due to its versatility.

We design a simulatability experiment to judge the quality of explanations. Here, we define quality as fidelity to the underlying model (Wiegreffe and Pinter, 2019; Jacovi and Goldberg, 2020) and information content that provides sufficient insight into the underlying model.

Our studies are performed using the Prolific crowdsourcing platform.[3] We randomly select a subset of dataset instances from the 676 HOTPOTQA validation instances with silver decompositions. Participants are paid at \$15/hour, and we qualify participants by first giving them a qualification question and verifying answers manually. We require participants to be located in the U.S. and to speak English as a first language. For each set of experiments, we source a distinct set of participants (no overlap) to avoid any bias in annotations that could occur from seeing past versions of the task or questions. For all experiments, we report Fleiss' κ (Fleiss, 1971) for binary or nominal data, and Kendall's τ (Kendall, 1938) for ordinal data.

For performance metrics, we report accuracy, F1, precision, recall, and Matthew's Correlation Coefficient (MCC; also known as the *phi* coefficient outside machine learning).

[3] https://www.prolific.co/

Model	Model Pred.	n	Sub-Q Accuracy
T5	Correct	617	**85.09**
	Incorrect	59	64.41
BART	Correct	597	**85.59**
	Incorrect	79	60.76

Table 3: Combined sub-question task performance, split by whether the model predicted the main question correctly or not.

3 Sub-question answering can distinguish incorrect and correct model predictions

We first investigate the extent to which performance on sub-question-answering is tied to performance on the main QA tasks. We split the validation set instances into two groups: those for which the model predicts the answer for the main question correctly, and those for which it does not. Results are presented in Table 3, which suggests sub-question accuracy is indicative of model performance, with a meaningful difference in sub-question accuracy observed between the instances which the model predicts correctly vs. those it does not.

4 Sub-question explanations allow humans to predict model behavior

Given the correlations between model's performance on main QA and sub-QA, we take a step forward to ask: can humans gain any useful insight from such correlations? We perform a simulatability experiment to measure how well the sub-question explanations can help humans predict model behaviors on the main HOTPOTQA task.

To this end, we design and conduct a human participant study to investigate crowd annotators' ability to make accurate predictions about model performance given question decompositions as explanations, following the protocol given in §2.5. We select a random 100-instance sample from the 671 HotPotQA validation instances, balanced such that the model predicts 50 instances correctly and 50 incorrectly, and perform the probing procedure described in §2.4 on the best-performing model (T5-Base), which results in tuples of the form (Q, C, A, Sub-Q_1, Sub-A_1, Sub-Q_2, Sub-A_2), where all answers A are predicted by the model.

Our goal is to observe how much the Sub-QA explanations help human annotators predict model behavior over a baseline that does not include these explanations, as well as investigate how the context (C) & the predicted answer (A) could potentially impact human's performance of diagnosing model errors. We design five different settings in which human participants are provided with different combinations of information. After reading the combination of information we present, the participants are asked to make their predictions about model's behavior on the main question (Q), i.e., whether or not the model will be able to correctly answer the given Q.

We recruit 50 participants on Prolific and split them into 5 batches, each of which contains 10 participants. Given the 100-instance sample, we split it into 5 batches of 20 questions each. We follow a Latin Square design, similarly to (Gonzalez and Søgaard, 2020), to ensure that each group of participants only sees each set of questions under one condition: (Q, A), (Q, A, Sub-Q), (Q, A, Sub-Q, Sub-A), (Q, Sub-Q, Sub-A), or (Q, C, Sub-Q, Sub-A), yet each condition is tested on both all 50 annotators and all 100 questions. This ensures that no bias in human predictions occurs due to having previously seen the questions and model predictions. Example of the UI that participants see is given in Figure 2. Finally, we collect their predictions and compute the performance scores using the actual main question's answer correctness as the ground truth.

Results are presented in Table 4. The average inter-annotator agreement is $\kappa = 0.24$. In order to ensure that human users are not simply performing the HotPotQA labelling task themselves, we validate this by first providing users with (Q, A) pairs, asking them "Do you think the answer to the given multi-hop question provided by the question-answering system is correct?". Because they are not given the context, C, this serves as a lower bound in quantifying any biases the participants may have about AI systems.

We apply the two-sided Mann-Whitney U test (Mann and Whitney, 1947) for statistical significance on accuracy numbers. Participant accuracy given (Q, A, Sub-Q, Sub-A) is statistically significantly different at $p = 0.01$ from all other settings, and results in substantially higher performance across all metrics except recall. This demonstrates that our proposed Sub-QA explanation method does help humans make more accurate predictions about model behavior on the main question (Q) than simply seeing model predictions (Q, A). We additionally validate that both sub-questions and sub-answers are important—when we ablate sub-answers, humans do poorly at the simulatability task given (Q, A, Sub-Q), resulting in no significant performance difference over (Q, A) pairs.

Having the answer (A) greatly improves the prediction performance, whereas the context (C) does not significantly impact human's prediction performance. Meanwhile, the proved feasibility of human's making accurate prediction about model behavior using Sub-QA explanations suggests a potential future direction for establishing an alternative for carrying out real annotation activities in order to diagnose QA system's error. The benefit of such alternative is obvious: humans will no longer have to conduct the question decomposition and perform the actual multi-hop reading comprehension by themselves. Instead, they may solely rely on or at least gain useful insights from their mental model about the QA system to save time and effort when trying to diagnose the error.

5 Quality of question decompositions matters

Prior work has shown that predictions from question decomposition models can improve task performance on HotPotQA when part of a larger modular system (Min et al., 2019; Perez et al., 2020; Khot et al., 2021), but qualitative inspection reveals a lack of quality in many cases. To investigate whether such sub-question-generation models can provide interpretability, we explore the effect of sub-question quality on utility of question decompositions as explanations in our probing setup. Namely, we conduct simulatability experiments and measure performance variation in humans' ability to guess model predictions based on

Setting	Metric				
	Acc.	F1	Precision	Recall	MCC
(Q, A)	$58.17_{1.55}$	$65.74_{1.63}$	$55.36_{1.59}$	$\mathbf{83.48}_{2.29}$	$19.15_{3.40}$
(Q, A, SUB-Q)	$56.57_{1.20}$	$62.94_{1.82}$	$53.78_{1.71}$	$79.32_{2.80}$	$15.41_{2.87}$
(Q, A, SUB-Q, SUB-A)	$\mathbf{63.50}^{*}_{1.39}$	$\mathbf{68.95}_{1.15}$	$\mathbf{60.82}_{1.46}$	$82.12_{1.60}$	$\mathbf{29.50}_{2.92}$
(Q, SUB-Q, SUB-A)	$53.07_{1.43}$	$61.25_{1.54}$	$52.49_{1.71}$	$76.88_{2.26}$	$8.29_{3.13}$
(Q, C, SUB-Q, SUB-A)	$57.00_{1.66}$	$64.61_{1.62}$	$54.82_{1.72}$	$80.37_{1.78}$	$14.79_{3.67}$

Table 4: Simulatability performance of human participants on 100 validation instances of HOTPOTQA given different input combinations. The majority baseline for accuracy is 50.00 since the dataset is fully balanced. All the statistics are computed by averaging across 50 participants, with standard errors included in subscripts. $*$: The setting's accuracy score distribution over 50 annotators is statistically significantly different from *all other methods* at $p = 0.01$ using two-sided Mann-Whitney U tests.

the quality of the SUB-QA explanations they received.

We use decomposition predictions from three trained question decomposers developed as part of larger modular QA systems in prior work: a) MODULARQA (Khot et al., 2021); b) One-to-N Unsupervised Sequence transduction (ONUS; Perez et al., 2020); and c) DECOMPRC (Min et al., 2019). MODULARQA is a next-question-prediction BART-LARGE model trained on the silver decompositions described in §2.2. ONUS is trained to decompose complex questions from the internet into simpler questions using supervision from noisy pseudo-decompositions. DECOMPRC is trained on a mix of supervision and heuristics to create sub-questions from the tokens in the original question, framing the task as span prediction. Examples of the question decompositions produced by each method are in Table 5. ONUS and DECOMPRC always produce two sub-questions; MODULARQA follows the form of SILVER and thus also results in 2 sub-questions per-instance once math operations are removed (§2.2).

We repeat the crowdsourcing process in §4, randomly sampling a subset of 30 correctly-predicted instances and 30 incorrectly-predicted instances from the 100 selected in §4. We probe the T5 model with SUB-Q_1 and SUB-Q_2 produced by each of the 4 sources: {SILVER, MODULARQA, ONUS, DECOMPRC}, and collect its SUB-A_1, SUB-A_2 responses. Tuples of (Q, A, SUB-Q_1, SUB-A_1, SUB-Q_2, SUB-A_2) are presented to 30 new annotators (who have not participated in previous experiments) following the setup in §4, where A represents the model's prediction.

Similar to §4, we perform a Latin Square design by equally splitting the participants and the questions into 3 batches, such that each participant group only observes each subset of questions under one experimental condition (either MODULARQA, ONUS, or DECOMPRC predictions). Annotator performance metrics at predicting answer correctness, averaged across all 30 participants, are presented in Table 6, along with annotator performance on the same subset given SILVER sub-questions. The average inter-annotator agreement is $\kappa = 0.29$.

We apply the two-sided Mann-Whitney U test (Mann and Whitney, 1947) for statistical significance on accuracy numbers. Human performance scores from the trained decomposers are all worse than the SILVER decomposer at a statistically-significantly different level ($p = 0.05$). This indicates that there are still notable gaps between the quality of SILVER's and other existing decomposers' SUB-QA explanations. ONUS question decompositions consistently provide the least explanatory power. Despite DECOMPRC's methodological simplicity, the explanatory power of its question decompositions is comparable to MODULARQA, though MODULARQA is the highest-performing predictive model overall. This is further supported by statistical significance results, which reveal that both MODULARQA and DECOMPRC are statistically-significantly different from ONUS, but not from one another ($p = 0.05$).

To further investigate the quality differences of different sources of question decompositions as measured by *human preferences*, we conduct an additional study where participants are asked to rank sources of SUB-QA explanations based on their quality. Specifically, we again recruit 30 new participants and each of them is asked to rank four decomposers' SUB-QA explanations for 30 random question samples in terms of three criteria: well-formedness, relatedness, and informativeness.

SILVER (Khot et al., 2021)
Sub-question 1: During what war was Pavillon du Butard occupied by the Prussians? **Sub-question 2:** What was the name that the French called the Franco-Prussian War?
MODULARQA (Khot et al., 2021)
Sub-question 1: During what war was the Pavillon du Butard occupied? **Sub-question 2:** What is the French name for the Franco-Prussian War?
ONUS (Perez et al., 2020)
Sub-question 1: What is the name the french give to the war? **Sub-question 2:** During which war did the prussians occupy the pavillon du butard?
DECOMPRC (Min et al., 2019)
Sub-question 1: which war during which the prussians occupied the pavillon **Sub-question 2:** what is the name the french give to Franco-Prussian War du butard?

Table 5: Examples of the question decompositions produced by {SILVER, MODULARQA, ONUS, DECOMPRC} for the question "What is the name the French give to the war during which the Prussians occupied the Pavillon du Butard?".

	Metric				
Decomposer	**Acc.**	**F1**	**Precision**	**Recall**	**MCC**
SILVER	**63.50**$^{*}_{1.39}$	**68.95**$_{1.15}$	**60.82**$_{1.46}$	**82.12**$_{1.60}$	**29.50**$_{2.92}$
MODULARQA	58.33$_{1.94}$	63.19$_{2.06}$	59.11$_{1.54}$	69.22$_{3.07}$	16.23$_{4.22}$
DECOMPRC	57.67$^{*}_{1.51}$	60.90$_{1.79}$	60.33$_{1.55}$	64.61$_{3.34}$	15.64$_{3.07}$
ONUS	53.11$_{1.53}$	52.80$_{2.70}$	55.96$_{1.53}$	54.13$_{4.05}$	6.07$_{3.22}$

Table 6: Simulatability performance of human participants on 60 validation instances of HOTPOTQA, where SUB-Q is provided by different question decomposers and SUB-A answers are obtained from our T5-BASE model. The majority baseline for accuracy is 50.00 since the dataset is fully balanced. All but SILVER statistics (copied from Table 4) are computed by averaging across 30 participants, with standard errors included in subscripts. $*$: The method's accuracy score distribution over 30 annotators is statistically significantly different from the method below it at $p = 0.05$ using a two-sided Mann-Whitney U test.

Example of the UI that participants see is given in Figure 3. Results are presented in Table 7. Inter-annotator agreement, as measured by Kendall's Tau (Kendall, 1938), is $\tau = 0.32$. SILVER decomposer is consistently preferred under all measurement criteria; MODULARQA is consistently second-best, followed by ONUS and DECOMPRC. This echoes results reported in Khot et al. (2021) (who only compared MODULARQA to DECOMPRC).

6 Related Work

Multiple prior works have concluded that question-answering as a *form* (Gardner et al., 2019) is a good choice for probing pretrained models (Roberts et al., 2020; Marasović et al., 2021). Roberts et al. (2020) fine-tune a model on a dataset of questions and answers, but claim this does not introduce new information to the model and only teaches form for effective QA probing. However, this claim is not well-supported, as fine-tuning removes any guarantees that the questions answered at test-time reflect information learned during pre-training alone, and is not zero- or few-shot. We avoid this by performing probing in a truly zero-shot manner (i.e., **we never fine-tune on sub-questions**). Additionally, the method of Roberts et al. (2020) does not probe for instance-level prediction explanations; the authors instead use a fixed set of questions on general topics. In our work, we use the instance-level explanations we obtain from probing with sub-questions to test whether these explanations give humans an accurate mental model of the system (Jacovi et al., 2022).

Most related to our work is that of Tang et al. (2021), who also investigate whether model architectures for multi-hop QA can answer single-hop questions. They find that there is a significant percentage of questions for which the model answers the main question correctly, but cannot correctly answer the multi-hop questions. However, because they use a model to produce question decompositions, their results may be confounded by errors or low quality of the questions themselves, which our work circumvents by using a silver source of

Decomposer	Well-formedness 1st	2nd	3rd	4th	Relatedness 1st	2nd	3rd	4th	Informativeness 1st	2nd	3rd	4th
SILVER	**50.2** $_{3.3}$	37.3 $_{2.9}$	6.4 $_{1.3}$	6.1 $_{1.2}$	**41.8** $_{3.0}$	**37.6** $_{2.7}$	11.5 $_{1.4}$	9.1 $_{1.3}$	**41.7** $_{3.0}$	37.4 $_{2.8}$	11.7 $_{1.3}$	9.2 $_{1.3}$
MODULARQA	35.5 $_{2.8}$	**45.1** $_{3.5}$	12.0 $_{1.9}$	7.4 $_{1.5}$	37.0 $_{2.5}$	36.7 $_{2.9}$	14.9 $_{1.6}$	11.4 $_{1.3}$	36.5 $_{2.5}$	**38.3** $_{2.9}$	14.2 $_{1.8}$	11.0 $_{1.4}$
ONUS	9.1 $_{1.3}$	13.2 $_{2.0}$	**54.9** $_{4.5}$	22.8 $_{2.2}$	14.1 $_{1.3}$	18.6 $_{1.6}$	**37.8** $_{3.6}$	29.5 $_{2.2}$	15.7 $_{1.5}$	17.0 $_{1.6}$	**40.1** $_{3.4}$	27.2 $_{2.0}$
DECOMPRC	5.2 $_{1.5}$	4.4 $_{1.0}$	26.7 $_{2.7}$	**63.7** $_{3.8}$	7.1 $_{1.3}$	7.1 $_{1.2}$	35.8 $_{2.4}$	**50.0** $_{3.2}$	6.1 $_{1.3}$	7.3 $_{1.2}$	34.0 $_{2.5}$	**52.6** $_{3.1}$

Table 7: Percentages (%) of the time each decomposer is listed in a ranking spot. Human participants rank all four question decomposers in terms of the well-formedness, relatedness, and informativeness of their corresponding questions and answers. Each annotator judges the same 30 instances, and results are averaged across 30 annotators. Subscripts indicate standard errors over 30 annotators.

sub-questions; we also investigate the effect of sub-question quality on the final results.

7 Conclusions

We have demonstrated the utility of question decomposition as an effective means to probe pretrained multi-hop question-answering models for supporting evidence. Through simulatability experiments, we show the effectiveness of this explanation form at allowing humans to predict model behavior, a sign that it helps humans to form an accurate *mental model* of the machine learning system. This ability to predict system performance occurs at the instance level instead of a sense of trust of the overall system, which can be important if the accuracy of the system is variable based on the question.

Our results indicate that explanations based on decompositional probes can be beneficial to users when the sub-questions are of reasonable quality. Our analyses indicate that existing decomposition systems, however, have considerable room for improvement. We can now look at the state of research in decomposition systems not only as to whether they improve multi-hop question answering, but whether they provide users with more calibrated trust.

The limitation is the need for higher-quality question decompositions. In future work, we hope to investigate and improve upon predictive models for question decomposition. We also plan to study balanced fine-tuning strategies that enable models fine-tuned on tasks of different formats, such as yes/no questions, to be probed.

References

Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and S. Kambhampati. 2019. Plan explanations as model reconciliation. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 258–266.

Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1036–1042, Brussels, Belgium. Association for Computational Linguistics.

Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It takes two to tango: Towards theory of ai's mind. *ArXiv*, abs/1704.00717.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. Question answering is a format; when is it useful?

Ana Valeria Gonzalez and Anders Søgaard. 2020. The reverse turing test for evaluating interpretability methods on unknown tasks. In *NeurIPS Workshop on Human And Machine in-the-Loop Evaluation and Learning Strategies*, volume 61, page 62.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, and Katja Filippova. 2022. Diagnosing ai explanation methods with folk concepts of behavior. *ArXiv*, abs/2201.11239.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

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

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2021. Text modular networks: Learning to decompose tasks in the language of existing models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1264–1279, Online. Association for Computational Linguistics.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. 2021. Few-shot self-rationalization with natural language prompts.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109, Florence, Italy. Association for Computational Linguistics.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8864–8880, Online. Association for Computational Linguistics.

Russell Perkins, Zahra Rezaei Khavas, and Paul Robinette. 2021. Trust calibration and trust respect: A method for building team cohesion in human robot teams. *ArXiv*, abs/2110.06809.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Yixuan Tang, Hwee Tou Ng, and Anthony Tung. 2021. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3244–3249, Online. Association for Computational Linguistics.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9073–9080.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

A Additional Details

We use Huggingface Datasets (Lhoest et al., 2021) and Huggingface Transformers (Wolf et al., 2020). Models are trained with a learning rate linearly decaying from $5E-5$, a batch size of 64, and default values for Adam (Kingma and Ba, 2015), gradient clipping, and dropout. We train for a maximum 200 epochs, performing early stopping on the validation loss with a patience of 10 epochs. All models are trained on a NVIDIA GeForce GTX 1080 GPU (8 GB memory) and on average take approximately 14 hours to train, converging in around 12 epochs. Input-output formatting is:

```
input_string = (f"question: {question}
    context: {passage}")
output_string = (f"{answer}")
```

The HOTPOTQA dataset has 90,447 train and 7,405 validation instances. In the HOTPOTQA leaderboard, there are two evaluation settings: distractor and full-wiki. In distractor, models are given 10 paragraphs where 2 of them are gold paragraphs needed to answer the question and the other 8 are "distractors". In the full-wiki setting, models are given the first paragraphs of all Wikipedia articles without the gold paragraphs specified. We do not submit to the leaderboard and thus cannot report test set performance, since we simplify the task and pass the 2 gold context paragraphs as input directly (§2.3) which does not align with either evaluation setting.

The Prolific interfaces for the human participant studies conducted in section 4 and section 5 are shown in Figure 2 and Figure 3, respectively.

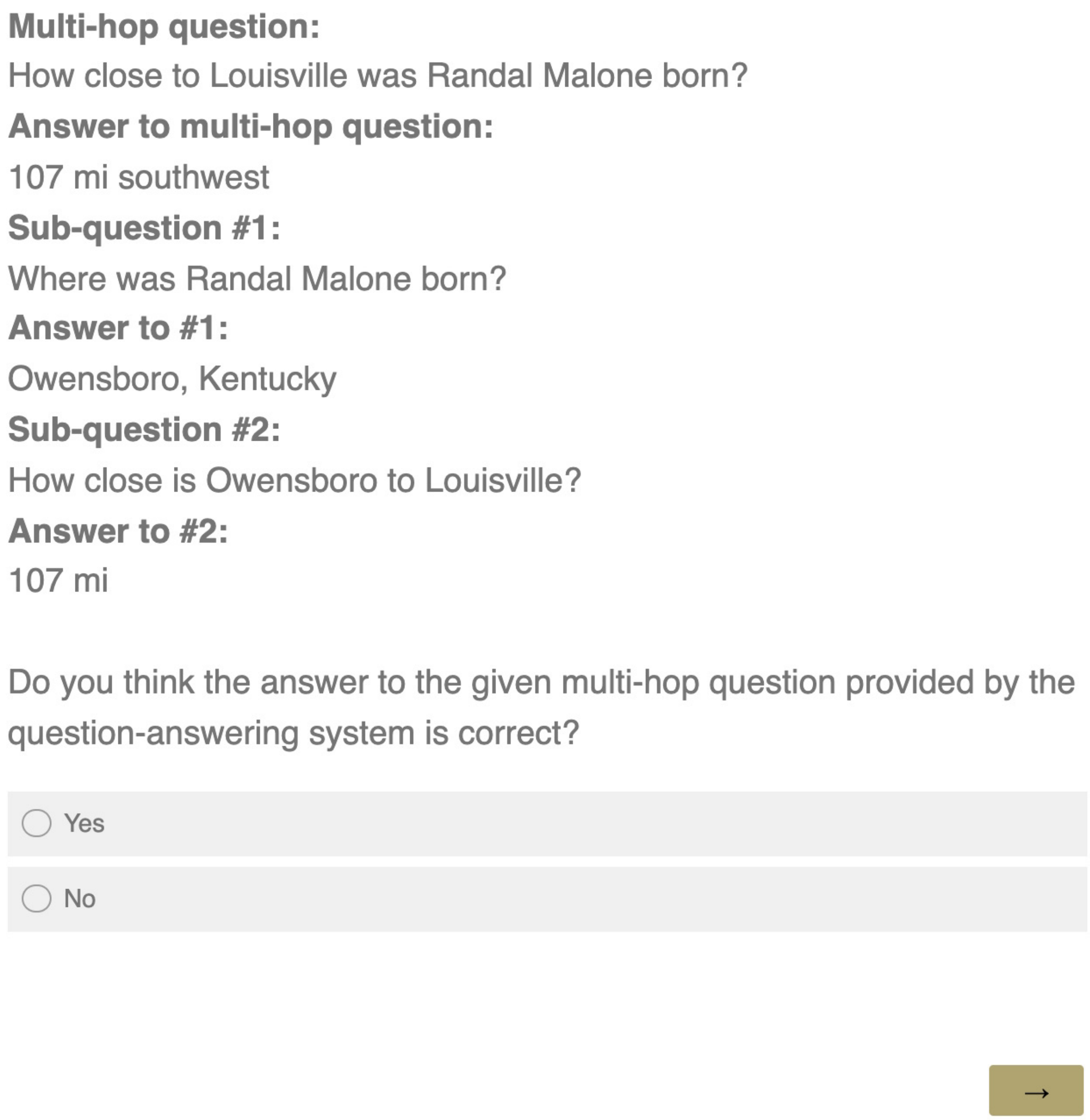

Figure 2: The Prolific interface for simulatability experiments in section 4.

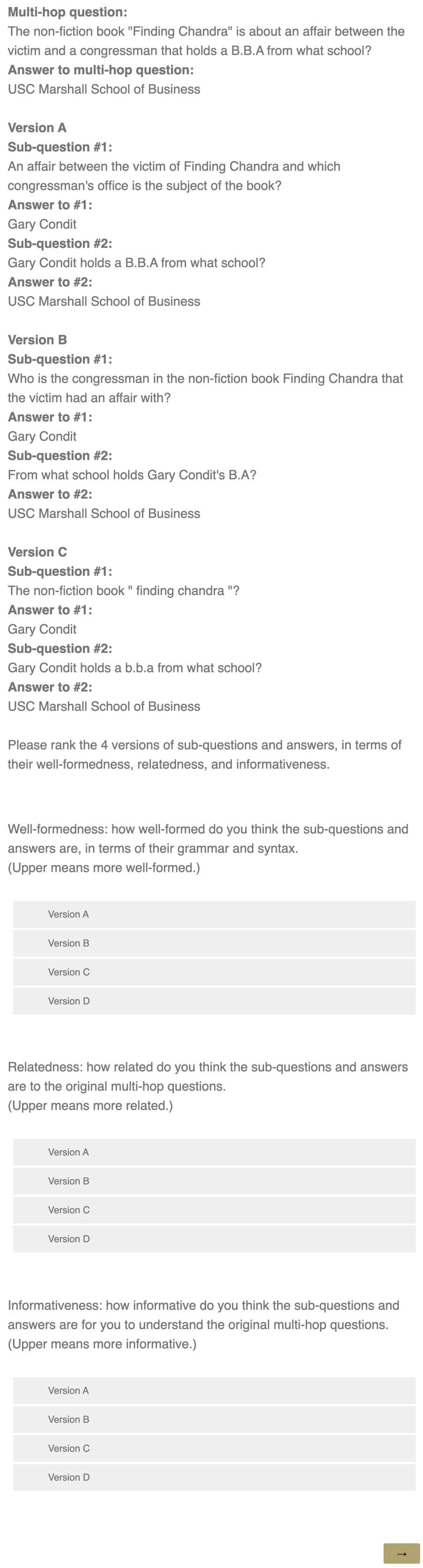

Figure 3: The Prolific interface for ranking experiments in section 5.