How Adversarial Can You Get Against LLMS?
Evaluation and Creation of Adversarial Examples in Question Answering

Anonymous ACL submission

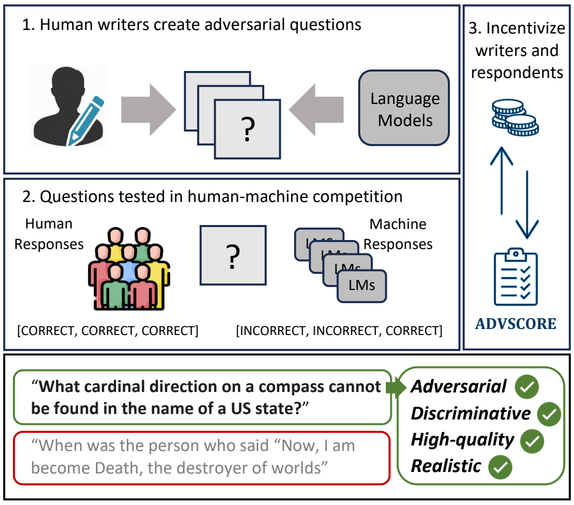
Abstract

Adversarial benchmarks validate model abilities by providing samples that fool models but not humans. We introduce an evaluation metric, ADVSCORE, that quantifies how adversarial and discriminative questions are. We then use ADVSCORE to create a pipeline that incentivizes writing good adversarial questions. We collect an adversarial QA dataset, ADVQA, from our pipeline's interface for eliciting human-authored adversarial examples. Questions in ADVQA surpass those in four challenging datasets across domains at not fooling humans but still fooling several language models, including GPT4. Additional analyses validate that ADVQA contains realistic and high-quality questions, based on difficulty estimates from 4890 human responses and responses from six models. Our evaluation pipeline is easily portable from QA to other domains.

1 Introduction

Language models are achieving near-perfect performance on many non-adversarial benchmarks (Bowman and Dahl, 2021a; Bowman, 2023), yet these systems fail to achieve comparable real-world performance (Ribeiro et al., 2020; Recht et al., 2019). Adversarial benchmarks help to fill this gap by identifying failures of strong models; however, adversarial examples must be challenging to capture model failures (Bowman and Dahl, 2021b).

One method of producing challenging examples is adversarial data collection, in which humans interact with a strong model in real time and produce examples that fool the models (Nie et al., 2020). Similarly, DADC[1] invites continuous human-model interaction (Kiela et al., 2021) as human authors write questions to probe the current models in iterative rounds (Wallace et al., 2022; Bartolo et al., 2020).


1. Human writers create adversarial questions
?
Language Models
2. Questions tested in human-machine competition
Human Responses
?
Machine Responses
LMs
[CORRECT, CORRECT, CORRECT]
[INCORRECT, INCORRECT, CORRECT]
3. Incentivize writers and respondents
ADVSCORE
Adversarial
Discriminative
High-quality
Realistic

"What cardinal direction on a compass cannot be found in the name of a US state?"

"When was the person who said "Now, I am become Death, the destroyer of worlds"

Figure 1: An overview of our new evaluation pipeline used to create ADVQA, our adversarial question dataset. Using ADVSCORE, it permits fine-grained evaluation of adversarial examples based on how *adversarial*, *discriminative*, *high-quality*, and *realistic* they are. For example, the question in the green box (from ADVQA) represents a good, adversarial question that meets our four criteria (by contrast, the question in the red box is neither high-quality nor realistic). Our pipeline incorporates a human-machine competition to collect real responses from each human and machine team and leverages item response theory for per-example quantitative assessment of human-authored questions.

However, Bowman and Dahl (2021b) argue that trivial artifacts in adversarial datasets obscure the abilities of the models, such that "benchmarks can be deceiving" (Kiela et al., 2021; Thrush et al., 2022). For example, the question "What popular rapper wrote a song about a woman in the 1990s?" may fool models due to its ambiguity, not because the models lack commonsense knowledge or reasoning ability.

We aim to aid the creation of good adversarial datasets that avoid trivial artifacts by introducing an evaluation pipeline (Figure 1) for creating adversarial datasets while filtering out subpar examples. We define the criteria for a *good adversarial question* as follows:

[1] https://dynabench.org/tasks/qa

Adversarial The questions should fool strong models while not fooling humans.

Discriminative The questions should distinguish between respondents with high and low skill levels (Boyd-Graber and Börschinger, 2020). If all models fail on all questions in a dataset, or all models succeed, the dataset does not help to distinguish whether some of those models are more robust to adversarial examples than others.

High-quality The questions should be adversarial for reasons that identify model weaknesses, such as the inability to compose clues or exclude redundant clues (Min et al., 2020, 2022). We base our criteria for what constitutes a useful adversarial tactic on the taxonomy of adversarial categories in Wallace et al. (2019).

Realistic The questions should be unambiguous and natural to humans, and represent potential issues that models could encounter in practice.

Our pipeline also introduces ADVSCORE, a metric to identify two of the four criteria: whether questions are adversarial and discriminative. In addition, we set other qualification criteria to ensure that the questions are *high-quality* and *realistic* (§ 6.2).

We run a human vs. computer QA competition, whose answer data is used by ADVSCORE. To account for the range of abilities in populations of the humans and models that answer questions, we adopt item response theory (Sedoc and Ungar, 2020; Lalor et al., 2019, 2016, IRT) to mathematically model respondents' answers (§ 2).

With ADVSCORE, each question is scored by quantifying per-question difficulty and discriminability, with IRT estimation of human vs. model responses. In our evaluation pipeline, this metric has two functions: 1) it incentivizes both human respondents and model designers to achieve the highest accuracy, and 2) it incentivizes authors to write better questions in our pipeline's interface (§ 3).

We use this incentive structure to collect a new adversarial question-answering dataset, ADVQA, via an interface designed to present the predictions of current LMs and retrieval models (§ 4). To confirm that our evaluation pipeline helps in crafting adversarial questions that meet the criteria, we compare ADVQA to other adversarial datasets and find that ADVQA contains the most questions that fool models while not fooling humans. We also validate ADVQA's scalability by crowdsourcing a second round of human responses for human evaluation of the questions (§ 7).

Our new evaluation pipeline and ADVSCORE metric contribute to the quantitative assessment of each question and creation of a new adversarial QA dataset, ADVQA. Our evaluation pipeline and metric center on QA datasets but easily generalize to other tasks (e.g., fact checking). Thus, in addition to our dataset, our core contributions include a reusable definition of adversarial benchmark quality and its evaluation mechanism.

2 Usage of IRT for AdvQA

A longstanding definition of an adversarial question is a question that a human can answer correctly but a computer cannot (Ribeiro et al., 2018). We suggest a more concrete definition that accounts for the range of abilities in populations that answer questions (Lord et al., 1968; Hopkins and May, 2013) based on IRT (Baker and Kim, 2004; Lalor et al., 2016). IRT allows a direct measurement of how *adversarial* and *discriminative* a question is.

IRT estimates latent variables (e.g., abilities) by modeling the associations between each subject and items (e.g., question difficulty and discriminability). The one-parameter logistic (1PL) model estimates the ability of a subject $i \in \mathcal{I}$ as β_i and the question difficulty $j \in \mathcal{J}$ as θ_j (Lalor et al., 2019). The higher the subject's skill is compared to the question's difficulty, the more likely the subject is to answer the question correctly. Taken together, this induces a probability $P_{ij}(r_{ij})$ that subject i will answer question j correctly, given that r_{ij} is a binary response of a subject i successfully answering question j (Martínez-Plumed et al., 2019):

$$P_{ij}(r_{ij} = 1 \,|\, \theta_j, \beta_i) = \frac{1}{1 + e^{-(\beta_i - \theta_j)}}. \qquad (1)$$

We also estimate question discriminability to encode how effectively the question rewards skill.[2] Thus, good questions have higher discriminability. We use the two-parameter logistic (2PL) model with a latent variable for discriminability γ_j (Baker and Kim, 2004):

$$P_{ij}(r_{ij} = 1 \,|\, \theta_j, \beta_i) = \frac{1}{1 + e^{-\gamma_j(\beta_i - \theta_j)}}. \qquad (2)$$

[2] Perfect discriminability means that any subject with positive difference between skill and difficulty will answer the question correctly.

To estimate the IRT parameters, we use variational inference (Jaakkola and Jordan, 1997), with Gaussian priors (see Appendix A.5).

3 New Evaluation Pipeline for AdvQA

In this section, we discuss how we design the IRT-initiated ADVSCORE to evaluate the collected questions for being *adversarial* and *discriminative* in our evaluation pipeline. We then discuss the competition setups in § 4 that aid in writing *high-quality* and *realistic* questions.

3.1 Designing metrics for *adversarial* questions

We first design a metric, ADVSCORE, that measures how *adversarial* and *discriminative* a question is, then provide incentives to writing teams based on this metric. We leverage the 2PL IRT equation (Eq. 2) to allow intuitive interpretation of how adversarial the given question set is. To this end, given an author a and set of questions Q_a, we calculate the margin between human (h) and model (m) team's probability of correctly answering the question. We then take the expectation value as

$$\mu_a = \frac{1}{|Q_a|}\mathbb{E}_j\Big[g(h,j) - g(m,j)\Big], \qquad (3)$$

where $g(d,j) = \max_i(\sigma(\beta_i^d - \theta_j^d))$, $d \in \{h, m\}$, and σ is a sigmoid function. Here, we take $\max_i$ to consider the most competent team's ability.[3] The positive value of μ_a implies that the most skilled human team achieves better accuracy than most skilled model team (*adversarial*), while a negative value implies the opposite.

Second, the best question set should include questions with the highest aggregate discriminability κ_a, meaning that they distinguish between high-skilled respondents and low-skilled respondents (here, we consider both human and model teams). Thus, we leverage 2PL IRT equation (Eq. 2) to estimate γ_j of each question:

$$\kappa_a = \frac{1}{|Q_a|}\sum_{j\in\mathcal{J}^{(a)}} \gamma_j. \qquad (4)$$

Third, to assess whether a question set is discriminative while remaining realistic, individual questions should be at a range of *human* difficulties: some questions should be easier for humans, and some should be harder (we avoid questions that every human can answer or only an expert can answer). Thus, we encourage question sets' human difficulty to have as large a median absolute deviation δ as possible:

$$\delta_a = \text{median}\left(\left|\theta_j^{(h)} - \text{median}_{j\in Q_a}\theta_j^{(h)}\right|\right) \qquad (5)$$

3.2 Why use IRT?

Although we could directly estimate these probabilities (i.e., human and model accuracy), we use real data of people answering questions to estimate a probability distribution of human vs. model responses because it encodes additional information about both the subject and the question. For infeasible or less discriminative questions, this probability will be dragged downward even some subjects might be lucky enough to guess the answer (Rodriguez et al., 2021). This gives a higher contribution to the overall score if there are humans who can consistently answer more of the dataset than computers. For example, even though the most skilled human team correctly answered the question, "what is the name of the first mosque in the world that was built by Prophet Muhammed during his hijrah from Mecca to Medina?", that team's probability of answering correctly was only 56%, slightly over random guess (more examples in Appendix A.4). These questions may be answerable to knowledgeable humans but not to those who are unfamiliar with the domain.

3.3 Adversarial competition incentives

To obtain human response data for estimation of IRT variables θ_j and γ_j, we hold two adversarial competitions: a QA competition in which models and human teams from the trivia community competed, and a team-based question-writing competition. The answer data collected from the question-writing competition is used by the ADVSCORE to reward incentives in the writing competition; the incentives are specifically designed to help create *adversarial* and *discriminative* questions.

We choose a winning respondent team (b^*) in the question answering competition by identifying the team with the highest skill β (the most correct responses):

$$b^* = \arg\max_b \beta_b. \qquad (6)$$

[3] We intentionally use difficulty values for the most skilled respondents in each team to ensure that the metric accounts for the most challenging questions and highest-performing models.

Then, to incentivize a writing team,[4] we score each question set by summing the human-model probability margin, discriminability, and divergence scores:

$$\text{ADVSCORE}_a = |Q_a|\frac{(\mu_a + \kappa_a + \delta_a)}{3}. \quad (7)$$

4 ADVQA Interface

We provide an adversarial writing interface as a human-AI collaborative tool for the adversarial writing competition, motivated by You and Lowd (2022)'s finding that human-AI collaboration strengthens adversarial attacks. Writers first choose the topic[5] they would like to write a question on (the *target answer*), then view the interface in Figure 2. The interface provides a set of widgets that help writers craft an adversarial question with real-time feedback. We focus on supplying the writers with the model interpretations, inspired by Wallace et al. (2019), so that they could continuously counteract the model response and make better edits.

4.1 Eliciting incorrect model predictions

The center of the interface provides the Wikipedia page for the target answer, which they use to write the question. While the author is writing, the retrieval widget (Figure 2, bottom left) and QA models widgets (right) are updated, drawing on the interface from Eisenschlos et al. (2021). Motivated by Feng et al. (2018), we embed the input perturbation inside the question writing widget (top left) to highlight which words trigger the model predictions. For example, changing "company" to a different token would be most likely to change the prediction to something other than the answer "Apple" in Figure 2.

Retrieval Systems Users receive real-time feedback on QA systems' performance on their questions via the interface's fine-tuned retrieval and reader model components. The retrieval system outputs (Figure 2, bottom left) are evidence that are used as contexts to elicit QA system predictions (right). Authors can rephrase questions to avoid retrieving information that is likely to make QA systems answer correctly. When a retrieval model answers incorrectly, it is tagged "`Fooled This Machine.`" We use lightweight sparse and neural retrieval models for writer feedback: a TF-IDF baseline and Dense Passage Retrieval (DPR, Karpukhin et al., 2020). To ensure that DPR predictions are diverse and up-to-date, we create a database that indexes each sentence in a set of Wikipedia pages (see Appendix C.1). We then use the RoBERTa-based FARMReader, which is fine-tuned on SQUAD (Rajpurkar et al., 2016), to read and sort the retrieved sentences from the two retrieval models by their relevance. If the target answer appears at the top of the retrieval widget, which means the author failed to fool the reader, they should revise the question so that FARMReader fails to extract the positive evidence for the QA systems.

LM-based QA Systems We enrich the model guidance by using both extractive and generative model answer predictions. For extractive QA, we use DistilBERT (fine-tuned on SQuAD), since its promptness and lightness facilitate rapid human-AI interaction. We also use T5 (Raffel et al., 2020) to answer the human-authored questions in a closed-book setting.[6]

4.2 Guiding writers to write *realistic* and *high-quality* questions

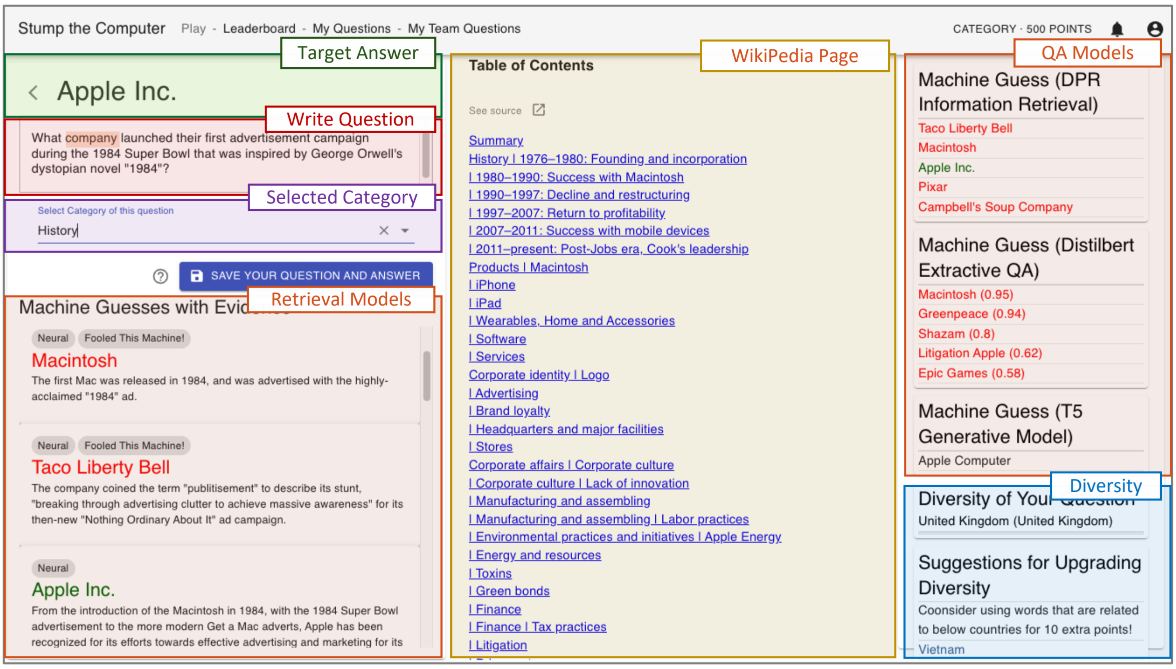
Inspired by Boyd-Graber and Börschinger (2020), we add another constraint in addition to generating adversarial questions in the interface: they must be *realistic*. To ensure the quality of the questions, we recruit experienced writers who are accustomed to trivia questions (more details in § 5). We filter questions that lack specificity and factuality, and avoid having many possible answers or highly subjective answers (details in Appendix 8).

Then, we perform an additional quality check by manually annotating what adversarial tactics the questions contain (details in Appendix A.9) in ADVQA. Inspired by Wallace et al. (2019), we add more tactics such as Novel Clues, Domain Expert Knowledge, and Location Misalignment. For example, a good question is *"What is the post-*

[4] We awarded a total $1100 worth of online gift cards after the competitions. The prizes were awarded to the first, second, and third winners, depending on each team's μ_a and α^*.

[5] Apart from fooling the models, we encourage topic diversity in the questions (Wang et al., 2020) by asking the authors to submit sets of questions with a fixed number of questions in each of nine categories: Art, Literature, Geography, History, Science, TV and Film, Music, Lifestyle, and Sports (Appendix A.7).

[6] The writing competition was held in Spring 2023, when DistilBERT and T5 were considered comparatively strong. We did not include CHATGPT because of its latency.



Apple Inc.

What company launched their first advertisement campaign during the 1984 Super Bowl that was inspired by George Orwell's dystopian novel "1984"?

Select Category of this question

History

Machine Guesses with Evidence

Macintosh

The first Mac was released in 1984, and was advertised with the highly-acclaimed "1984" ad.

Taco Liberty Bell

The company coined the term "publitisement" to describe its stunt, "breaking through advertising clutter to achieve massive awareness" for its then-new "Nothing Ordinary About It" ad campaign.

Apple Inc.

From the introduction of the Macintosh in 1984, with the 1984 Super Bowl advertisement to the more modern Get a Mac adverts, Apple has been recognized for its efforts towards effective advertising and marketing for its

Table of Contents

Summary
History | 1976–1980: Founding and incorporation
| 1980–1990: Success with Macintosh
| 1990–1997: Decline and restructuring
| 1997–2007: Return to profitability
| 2007–2011: Success with mobile devices
| 2011–present: Post-Jobs era, Cook's leadership
Products | Macintosh
| iPhone
| iPad
| Wearables, Home and Accessories
| Software
| Services
Corporate identity | Logo
| Advertising
| Brand loyalty
| Headquarters and major facilities
| Stores
Corporate affairs | Corporate culture
| Corporate culture | Lack of innovation
| Manufacturing and assembling
| Manufacturing and assembling | Labor practices
| Environmental practices and initiatives | Apple Energy
| Energy and resources
| Toxins
| Green bonds
| Finance
| Finance | Tax practices
| Litigation

Machine Guess (DPR Information Retrieval)

Taco Liberty Bell
Macintosh
Apple Inc.
Pixar
Campbell's Soup Company

Machine Guess (Distilbert Extractive QA)

Macintosh (0.95)
Greenpeace (0.94)
Shazam (0.8)
Litigation Apple (0.62)
Epic Games (0.58)

Machine Guess (T5 Generative Model)

Apple Computer

Diversity of Your Question

United Kingdom (United Kingdom)

Suggestions for Upgrading Diversity

Coonsider using words that are related to below countries for 10 extra points!

Vietnam

Figure 2: As the target answer to the question should be "Apple Inc," the interface is updated with answers from retrieval models with the most relevant sentence and from LMs (e.g., Distilbert, T5). Also, the highlights are updated by the input perturbation technique.

apocalyptic science fiction action film directed by a Korean director, but not by the director of Parasite, about the class struggles of passengers on a train attempting to survive their journey?": it misleads the model with a multistep-reasoning adversarial tactic, while remaining specific. With such instructions and interface, we guide the authors to write adversarial questions; in the next section, we discuss the details of competition results.

5 Adversarial Competitions

We collect 399 adversarial questions through the interface in the writing competition, recruiting 5 writer volunteers. Then, we hire an expert editor to scrutinize the human-authored questions for grammar errors or poor quality (§6.2).

In the answering competition, we invite eight human teams (composed of 3-4 people) and four model respondent teams (DPR, T5, DISTILBERT,[7] and CHATGPT). We use subsets of 30 questions for each match of our question answering competition. Here, the questions are qualitatively checked once more; as the players hear and judge the written questions, they note incoherent or unnatural questions and request to exclude them from packets. This process makes for an additional quality check on our dataset, resulting in 184 questions and 790 human responses.

6 ADVQA Evaluation

We evaluate ADVQA based on our definition of a good, adversarial question (§1) and verify that ADVSCORE helps in creating such adversarial questions.

6.1 Evaluating ADVQA with ADVSCORE

Are the questions *adversarial*? We assign the difficulty values by fitting a 2PL model that learns the latent variables θ_j and γ_j, estimated with variational inference.[8] We convert each respondent's free-form answer into a binary label (1 if correct, 0 if incorrect).

We run the model individually on the human and model responses to elicit the difficulty levels of the questions. First, we group the questions on the kind of respondents they fooled based on the human difficulty θ^h and model difficulty θ^m: "fooled both," "fooled only models," "fooled only humans," or "easy for both." Given a 2PL model assuming $\theta^h \sim \mathcal{N}(0, \rho)$, $\theta^m \sim \mathcal{N}(0, \rho)$, if both θ^h and θ^m exceed 0, we consider the question to

[7] Both finetuned on SQUAD (Rajpurkar et al., 2018)

[8] We used the Python library, https://pypi.org/project/py-irt/, a Python library for IRT.

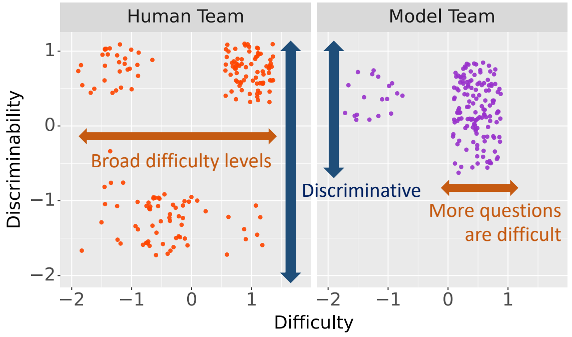


Figure 3: As intended, ADVQA questions contain all levels of difficulty for human respondents, while model teams regard most questions to be difficult. Moreover, both model and human teams' responses suggest that most questions are *discriminative*. Each team's discriminability and difficulty values are from individually fitted 2PL models.

Adversarial Tactics	Percentage
Composing Seen Clues	28.0%
Domain Expert Knowledge	2.1%
Location Misalignment	6.2%
Logic & Calculation	6.2%
Multi-step Reasoning	12.3.%
Negation	0.6%
Novel Clues	37.7%
Temporal Misalignment	2.7%
Commonsense Knowledge	1.4%
Crosslingual	2.7%
Failed to fool LM	20 %

Table 1: We analyze the questions that fooled GPT models. The most frequently used adversarial tactics are *Composing seen clues* and *Novel Clues*.

have "fooled both." If model difficulty is positive while human difficulty is not, we regard the question to have "fooled only models." After grouping the questions, we compute the margins between the human difficulty and model difficulty for each group of questions following Equation 3 in Section 3.1. 76% of the questions in ADVQA fool the models: 36% fool only the models and 40% fool both humans and models (Table 2).

Are the questions *discriminative*? We also analyze the correlation between the difficulty θ_j and discriminability γ_j in our questions (Equation 4). Figure 3 indicates that most questions were discriminative for humans, and that questions were at a range of human difficulties. For most models, the questions were difficult and discriminative. From the analyses, we infer that our pipeline not only helps create *adversarial* questions but also *discriminative* questions for models.

6.2 Are the questions *high-quality* and *realistic*?

After collection, we validate that questions are *realistic* in several ways. First, we note that there was an improvement in question quality when authored by trivia writers acquainted with "trivia norms" (Boyd-Graber and Börschinger, 2020; Rodriguez and Boyd-Graber, 2021) than the college students. 57% of questions written by trivia experts were labeled as acceptable by the trivia editor, compared to 38% of those written by college students. A common failure mode is writing vague or subjective questions, e.g., "What video game movie featuring one of the *most popular* and *well-known* icons in video games stars Chris Pratt and Jack Black?" Second, the expert trivia editor that we hired scrutinizes the human-authored questions for grammar errors or poor quality (see Appendix A.8). Third, our human vs. computer competition provides an additional quality check, as human respondents flag potential issues while answering questions. If they think a question is unnatural or ambiguous, we exclude it from our final dataset.

Moreover, to check for *high-quality* adversarial questions, we manually annotate what adversarial tactics the ADVQA questions contain (details in § 4.2). Among the 80% of ADVQA questions that fooled GPT-3.5, the questions most often fooled models by requiring abilities like *Composing seen clues* and excluding *novel clues* that are redundant to answer the question (Table 1). These results suggest that our adversarial questions are in *high-quality* and recognize the current hurdles that LMS face when deployed in practice.

7 Comparisons to Previous Benchmarks

We compare ADVQA to previous benchmarks to validate that our pipeline is effective for producing quality adversarial questions.

7.1 How *adversarial* and *discriminative* is ADVQA?

We perform an analysis to compare the proportions of *adversarial* questions in ADVQA to previous benchmarks. Because most adversarial benchmarks do not collect human responses (and thus do not verify that humans are able to answer the questions), we compare with two benchmarks that

Fooled Subjects	Proportion of adversarial questions (%)			
	AdvQA	FM2	Trickme	Bamboogle
Easy for both	13%	36%	60%	30%
Only humans	9%	20%	16%	8%
Fooled both	40%	21%	3%	30%
Only models	**36%**	22%	21%	33%
Rank	**1**	3	4	2

Table 2: We perform comparative analysis for our dataset. AdvQA placed first in terms of having the highest proportion of questions that fooled models but not humans (and the most questions that fooled models overall), meaning that our adversarial dataset has the highest proportion of adversarial samples.

do contain human responses, Trickme (Wallace et al., 2019) and FoolMeTwice (FM2) (Eisenschlos et al., 2021). FoolMeTwice tricks the model using entailment pairs; we use human responses from a user study by Si et al. (2023). Trickme fools QA systems using 99 pyramidal questions.[9] We also compare with another popular benchmark, Bamboogle, (Press et al., 2022) which contains questions that elicit incorrect answers when using Google's search engine. As Bamboogle did not have such responses available, we recruited trivia expert respondents to collect human responses to 125 questions (see also § 7.2). For model respondents, we use Llama-2-7b-chat (Touvron et al., 2023), GPT-4 (Achiam et al., 2023), and the neural retrieval model DPR (also used for our interface: see Section C.1) for our analysis.

We follow a procedure similar to §6.1 to examine the proportion of adversarial questions that "fooled only models" and "fooled both humans and models" for each dataset (Table 2). For both categories, AdvQA placed first: 36% of its questions fool only the models (+3% over the next-best dataset); and 76% of its questions fool models overall (+13% over the next-best dataset).

Moreover, we comparatively measure how adversarial AdvQA is using Equation 3 (Table 3). AdvQA had the highest μ and AdvScore among four datasets, indicating its questions were most adversarial. Bamboogle, FM2, Trickme datasets had near-zero μ values, meaning that both teams have balanced probabilities: questions are not adversarial. The negative AdvScores of FM2 and Trickme infer that estimated discriminability values do not contribute AdvScore: the questions are not likely to be discriminative.

Adversarial Datasets (a)	μ	AdvScore
AdvQA (Offline)	**0.89**	**33.83**
AdvQA (Online)	0.84	31.11
Bamboogle	0.004	25.09
FM2	0.005	-13.89
Trickme	0.004	-29.37

Table 3: μ of AdvQA (Offline), the expected margin between model and human probability in AdvQA, was higher than that of all other datasets tested. Also, AdvScore of AdvQA (Offline) had the highest value. Both measures indicate that AdvQA contains questions that are most *adversarial*. We validate the generalization of AdvQA's adversarial-ness by observing AdvQA (Online)'s next highest value in terms of μ and AdvScore.

7.2 Verifying generalization via human evaluation

To ensure that our unique in-person human collection generalizes to more conventional data collection methods, we collect a second set of answers for AdvQA online, alongside the answers to Bamboogle questions. We crowdsource 4100 responses from 165 members of the trivia community on 184 questions of AdvQA. Table 4 gives examples of questions from AdvQA that satisfy the criteria of being a good adversarial benchmark, contrasted by Bamboogle. After fitting 2PL models on human and model answers, we compare the μ_a and AdvScore$_a$ for each dataset, including offline and online responses for AdvQA (Table 3). μ of AdvQA (Offline) was higher than that of all other datasets tested. Also, AdvScore of AdvQA (Offline) had the highest value. Both measures indicate that AdvQA contains questions that are most *adversarial*. We validate the generalization of AdvQA's adversarial-ness by observing AdvQA (Online)'s next highest value in terms of μ and AdvScore.

[9] The human buzz points were used as model input to garner model predictions.

Bamboogle Question	**Answer**	**Human**	**GPT4**	**Human Remarks**
Who was the father of the father of computer science?	Julius Mathison Turing	Horatio Hornblower	Charles Babbage	"There are multiple fathers of computer science: Babbage, Zues, Turing"
What is the highest elevation (in meters) of the second largest island in the world?	4,884 m	3,500 m	5,030 m	"No human would no this"
AdvQA Question	**Answer**	**Human**	**GPT4**	**Grounding**
Who is the president of the country represented by the second letter in the acronym BRICS, which refers to countries with emerging economies?	Vladimir Putin	Putin	Russia	GPT-4 did not understand the question, which requires *composing clues* and *multi-hop reasoning*
What cardinal direction on a compass cannot be found in the name of a US state?	East	East	Wicked	GPT-4 hallucinated answer to question that requires *location aligning ability*

Table 4: Examples of Bamboogle questions with (unedited) human remarks indicating that they are not *adversarial* (e.g., not easy for humans) or *realistic*. On the other hand, examples of AdvQA questions are *high-quality* and follow adversarial tactics such as *composing seen clues*.

8 Related Work

Recently, the NLP community has questioned whether models trained on benchmarks learn to solve tasks in robust and generalizable ways (Ribeiro et al., 2020; Bartolo et al., 2021; Nie et al., 2018; Gururangan et al., 2018). Current systems are prone to error under plausible domain shifts due to lack of generalization (Kaushik et al., 2021). An alternative is to provide more challenging benchmarks that require a stronger form of generalization and diversity (Rychalska et al., 2019). For example, Ma et al. (2021) and Kiela et al. (2021) have collected data within dynamic adversarial generation frameworks in which humans create examples while interacting with the model.

However, Tedeschi et al. (2023) postulate that the abilities of many "superhuman" models may be overestimated due to poorly annotated datasets and biases embedded in the evaluation process (e.g., fixed test sets). Our adversarial dataset creation framework could not only help the experts to create the next generation of data, but also systematically probe models to understand their capabilities (Bowman, 2023; Yuan et al., 2023).

Turning to dynamic adversarial generation for QA, Bartolo et al. (2021) uses a synthetic generation method to create human adversaries. Sheng et al. (2021) introduces a benchmark in which the humans interact with a visual QA model, and write an adversarial question for each of a set of images. Wallace et al. (2019) and Eisenschlos et al. (2021) both use dynamic incentive mechanisms to create adversarial questions. To remedy the issue that current evaluation treats each model independently rather than considering relative differences, Lalor et al. (2019) introduces an IRT-based ranking method. Rodriguez et al. (2021) also considers this issue by redesigning the leaderboard framework with a Bayesian leaderboard model where latent subject skill and latent item difficulty predict correct responses.

9 Conclusion and Future Work

We introduced an evaluation pipeline for adversarial questions based on a new metric, ADVSCORE. We used our pipeline to construct ADVQA, a dataset with adversarial, discriminative, high-quality, and realistic questions. We validated that ADVQA has more adversarial questions than previous adversarial datasets and performed several quality checks to ensure that the questions are also realistic and high-quality; it can be used as test sets to stress-test model abilities. Though ADVQA is not large, such highly curated datasets are effective at revealing model weaknesses (Vania et al., 2021; Press et al., 2022; Kim et al., 2022).

Moreover, ADVQA can also serve as a seed dataset for generating similar questions via LMs, serving as an essential resource for harvesting a larger adversarial dataset. In addition, as our evaluation pipeline is not restricted to QA tasks, we contribute it as a way to assess adversarial examples in other domains. For future work, we plan to develop an evaluation pipeline with a reward system that not only compensates for good adversarial questions but also incentivizes those who help in good calibration of the trained model, allowing better interpretation where the models are fooled.

10 Limitations

Though we empirically validate the quality of our questions, a remaining concern is that because the human authors are given specific instructions to create adversarial examples, they may create task-diagnostic examples that focus on specific types of adversarial-ness (Bowman and Dahl, 2021a). For example, when someone learns that reasoning ability is missing from an LM, they may repeatedly create examples that require reasoning ability. This could result in data that is oriented towards widely known problems of LMs instead of uncovering new patterns for an adversarial benchmark(Kaushik et al., 2021). Thus, it will be worthwhile to specifically provide them with the taxonomy of adversarial tactics and require that they write questions within such taxonomy.

In addition, the abilities of the retrieval models in the interface are dependent on the retrieval corpora we used. For future uses of our pipeline, we suggest updating the retrieval corpora periodically to address this limitation and allow for rich model interpretation when writing adversarial questions.

Although we attempted to gather questions that were demographically diverse, we did not observe any significant diversity in the country distribution or the written question (Appendix C.2). We plan to improve our interface's "Diversity" widget by linking more named entities (NER) to different countries and thereby translating written sentences that include NERs into country representations.

11 Ethical Considerations

Our study was pre-monitored by an official IRB review board to protect the participants' privacy rights. The identity characteristics of the participants were self-identified by the workers by completing the task.

Before completing the task, we display consent forms for the workers to agree that their answers would be used for academic purposes. They were invited to participate in the writing and answering task for entertainment and academic purposes.

We emphasize the scale and the impact of our research in that it provides the resource and an evaluation metric, not constrained to QA, to resolve the current hallucinations and artifacts in NLP datasets.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Frank B Baker and Seock-Ho Kim. 2004. *Item response theory: Parameter estimation techniques*. CRC press.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel R Bowman. 2023. Eight things to know about large language models. *arXiv e-prints*, pages arXiv–2304.

Samuel R. Bowman and George Dahl. 2021a. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021b. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435, Online. Association for Computational Linguistics.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Mark Hopkins and Jonathan May. 2013. Models of translation competitions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1424, Sofia, Bulgaria. Association for Computational Linguistics.

Tommi S. Jaakkola and Michael I. Jordan. 1997. A variational approach to Bayesian logistic regression models and their extensions. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, volume R1 of *Proceedings of Machine Learning Research*, pages 283–294. PMLR. Reissued by PMLR on 30 March 2021.

Ken Jennings. 2007. *Brainiac: adventures in the curious, competitive, compulsive world of trivia buffs*. Villard.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih. 2021. On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study. 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 6618–6633, Online. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Najoung Kim, Phu Mon Htut, Samuel R Bowman, and Jackson Petty. 2022. QA^2: Question answering with questionable assumptions. *arXiv preprint arXiv:2212.10003*.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

John P Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. 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 4249–4259.

Adam D Lelkes, Vinh Q Tran, and Cong Yu. 2021. Quiz-style question generation for news stories. In *Proceedings of the Web Conference 2021*, pages 2501–2511.

Frederic M Lord, Meivin R Novick, and Allan Birnbaum. 1968. Statistical theories of mental test scores. 1968. *Reading: Addison-Wesley*.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Yu Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. In *Neural Information Processing Systems*.

Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. 2019. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial intelligence*, 271:18–42.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Sewon Min, Luke Zettlemoyer, Hannaneh Hajishirzi, et al. 2022. Crepe: Open-domain question answering with false presuppositions. *arXiv e-prints*, pages arXiv–2211.

Prathiba Natesan, Ratna Nandakumar, Tom Minka, and Jonathan D Rubright. 2016. Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2018. Analyzing compositionality-sensitivity of nli models. *ArXiv*, abs/1811.07033.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

John K Pollard. 2006. Student reflection using a web-based quiz. In *2006 7th International Conference on Information Technology Based Higher Education and Training*, pages 871–874. IEEE.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

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. *J. Mach. Learn. Res.*, 21(1).

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.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 856–865.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? 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 4486–4503, Online. Association for Computational Linguistics.

Pedro Rodriguez and Jordan Boyd-Graber. 2021. Evaluation paradigms in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9630–9642, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan L. Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. *CoRR*, abs/1904.04792.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10).

Barbara Rychalska, Dominika Basaj, Alicja Gosiewska, and Przemysław Biecek. 2019. Models in the wild: On corruption robustness of neural nlp systems. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III*, page 235–247, Berlin, Heidelberg. Springer-Verlag.

Joo Sedoc and Lyle Ungar. 2020. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, Online. Association for Computational Linguistics.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering. *CoRR*, abs/2106.02280.

Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2023. Large language models help humans verify truthfulness–except when they are convincingly wrong. *arXiv preprint arXiv:2310.12558*.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada. Association for Computational Linguistics.

Tristan Thrush, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gaviria Rojas, Peter Mattson, Adina Williams, and Douwe Kiela. 2022. Dynatask: A framework for creating dynamic AI benchmark tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 174–181, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. 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 1141–1158, Online. Association for Computational Linguistics.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. 2022. Analyzing dynamic adversarial training data in the limit. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 202–217.

Zhen Wang, Siwei Rao, Jie Zhang, Zhen Qin, Guangjian Tian, and Jun Wang. 2020. Diversify question generation with continuous content selectors and question type modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2134–2143, Online. Association for Computational Linguistics.

Wencong You and Daniel Lowd. 2022. Towards stronger adversarial baselines through human-AI collaboration. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 11–21, Dublin, Ireland. Association for Computational Linguistics.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

Quan Yuan, Mehran Kazemi, Xin Xu, Isaac Noble, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. Tasklama: Probing the complex task understanding of language models. *arXiv preprint arXiv:2308.15299*.

A Details on Dataset Creation

A.1 Goals in Dynamic QA Generation

When tasking human authors with adversarial writing of questions, Wallace et al. (2019) emphasizes the importance of “who” the authors should be: *talented and eager* question writers with *specific goals*; they should aim to generate questions that stump computers but seem normal enough for humans to answer. To make this work, they recruit members of the quizbowl community, who have deep trivia knowledge and craft question for quizbowl tournaments (Jennings, 2007). However, their challenge was to convey what is "normal" to authors and stimulate examples that can elucidate the weaknesses of QA models.

A.2 Merging Trivia Question Generation and Dynamic Adversarial Generation Process

Many QA datasets are now too easy for modern models as models have become more powerful (Rogers et al., 2023). However, even these easy QA datasets have serious data flaws (Min et al., 2020; Yu et al., 2023), which suggests that creating question-answer pairs is a very challenging task. This is also a norm for questions written for human players, where more than 100,000 questions are produced annually. To create effective and challenging enough questions, the professional experts (e.g., writing staff) take a rigorous editing pass on the questions to decide whether they are adequate enough to guarantee players a fair game (Lelkes et al., 2021; Pollard, 2006). They follow strict guidelines to be selected to be used in the quiz matches. We propose to merge the above pipelines to help improve data creation for robust QA models by adding an editing step to ensure that grammatical errors and nonfactual questions (following the norms of Trivia questions) do not exist in the pool.

A.3 Competition Details

We scrutinize the raw statistics of the questions that fooled the models but not humans. The number of questions that fooled *some* humans and *all* models was the highest (Table 5).

A.4 Details on errors in using raw scores in question answering competition

A.5 Variational Inference for IRT models

To discover the IRT parameters that best explain the whole data, difficulty θ_j and discriminabilityγ_j, we turn to variational inference for the full generative

Number of Questions fooled			
		Model	
		All	Some
Human	All	73	8
	Some	**90**	13

Table 5: The number of questions that fooled *some* humans and *all* models was the highest (90 questions), followed by questions that fool both humans and models. This indicates that ADVQA questions are difficult for models, and typically fool models while not fooling humans, fulfilling our goal of being adversarial.

process, an approximation method for intractable posterior distribution in Bayesian inference (Natesan et al., 2016; Lalor et al., 2019). The parameters θ and β follow Gaussian prior distributions and make inferences through joint posterior distribution $\pi(\theta, \beta|Y)$ (Natesan et al., 2016).

A.6 Displaying writer incentives

To encourage competition and allowsauthors to monitor their progress, authors can monitor how many questions they wrote per category and their diversity level on the `Writer Leaderboard` (Appendix C.3). Once the authors finish writing the questions, the `Machine Leaderboard` updates whether their questions stumped CHATGPT.

A.7 Topic Categories of Questions

We ask the question writers to tag their questions with the categories below. With reference to specific categories and examples, we encourage them to be as creative and diverse as possible when authoring the questions. In the interface, they can monitor how many questions they wrote per category. They are required to submit packets with a specific amount of questions in each category.

A.8 Question Type Annotation

In Table 8, we list the problematic question types that we ask the annotators to annotate. These are illustrated with descriptions and examples to help them better understand each question.

A.9 Adversarial Tactic Annotation

In Table 9, we list adversarial types (techniques) to determine how each question is using them to stump the models. The annotators are given the description and examples to better understand the reasons why the models may have been stumped. They are expected to tag the examples with the model prediction and question.

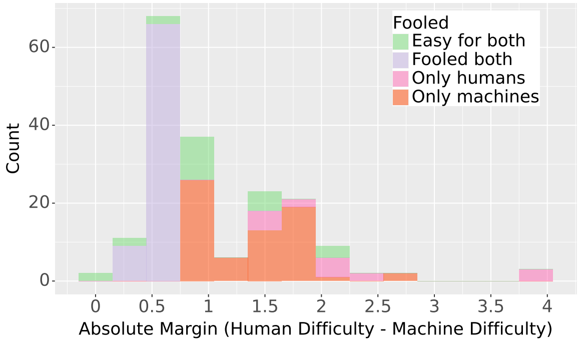


Figure 4: Our IRT analysis exposes that the samples that fooled only the models had the highest margin among other categories (e.g., fooled both or only humans).

A.10 Question Examples Annotated with Question and Adversarial Types

Table 10 shows question examples that are annotated with question and adversarial types. The highlights in the question correspond to either adversarial types or question types that are highlighted with the same color.

B Details on Dataset Analysis

B.1 Adversarial Frequency by Difficulty Margin

Our IRT analysis exposes that the samples that fooled only the models had the highest margin among other categories (e.g., fooled both or only humans).

B.2 Correlation between Adversarial Types and Discriminability

We scrutinize what kind of adversarial tactics were used by writers to stump LLMs and evaluate if they are "good" or "bad". To understand *how* they are bad, we examine if there is any correlation between the adversarial-ness the question has and how *good* they are. Figure 5 and 6 shows that *Temporal Misalignment*, *Composing Seen Clues*, *Domain Expert Knowledge*, and *Novel Clues* are used more frequently in questions with high discriminability. On the other hand, *Multistep Reasoning*, *Domain Expert Knowledge*, and *Logic & Calculation* are used less in questions with high discriminability.

B.3 Examples sorted by Difficulty Score

In Table 11 and 12, we demonstrate examples sorted by the learned variables difficulty (θ) from IRT model. The examples with the highest variable value is ranked 1.

Question	Gold Answer	Human Answer	Probability $\sigma(\beta_i - \theta_j)$
What phrase is common to the title of novel featuring a fictional Nat King Cole recording, a Gene Autry film and song, and an I-95 attraction between the Carolinas?	South of the Border	Correct	0.57
In which novel, written by an author who was originally a botanist and born in Cuba, features a fictitious conversation between a merchant who travelled a road that was known by a smooth natural material and an emperor who loved to write Chinese poetry, both of which are actual people in history?	Invisible Cities	Correct	0.55
What is the name of the first mosque in the world that was built by Prophet Muhammed (s.a.w) during his hijrah from Mecca to Medina?	Quba Masjid	Correct	0.56

Table 6: While the most skillful human team answered all three questions correctly, the estimated probability of the human teams answering the question correctly when compared to their ability was low (50%). This infers that the human accuracy does not necessarily translate to answering ability or question difficulty measurement, which obscures the measuring the the question's adversarial-ness.

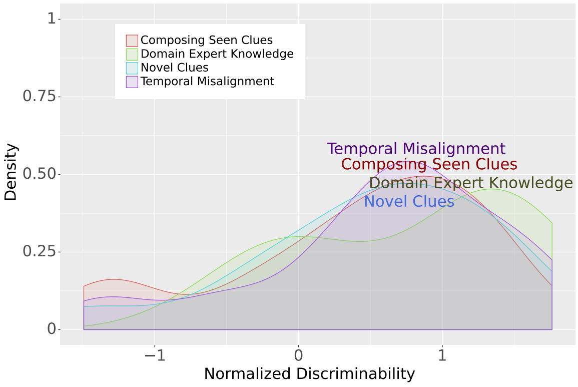


Figure 5: The adversarial techniques *Temporal Misalignment*, *Composing Seen Clues*, *Domain Expert Knowledge*, and *Novel Clues* are used more frequently in questions with high discriminability.

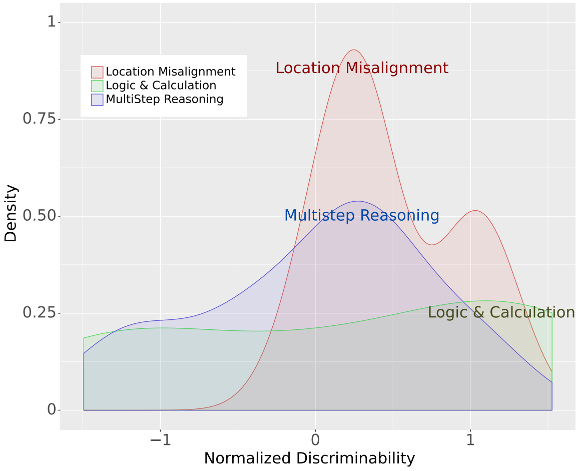


Figure 6: The adversarial techniques *Location Alignment*, *Multistep Reasoning*, *Domain Expert Knowledge*, and *Logic & Calculation* are used less in questions with high discriminability.

B.4 Explanation Examples from Retrieval Models and CHATGPT

In Table 13, we demonstrate the explanations from retrieval models and CHATGPT models to deeply analyze how explanations from retrieval model may help stump the CHATGPT.

C Interface Details

C.1 Retrieval System Details

To ensure that the retrieval results help in obtaining up-to-date information for the writers, we created the database for Wikipedia pages and DPR training data. DPR retrieves the most relevant sentence from a database that consists of the Top 1000 popular Wikipedia pages[10] from 2021 to 2022. DPR is finetuned with the 2018 and 2021 QANTA datasets (Rodriguez et al., 2019). For training, we used the questions and gold evidence as positive samples, and sentences from pages that are two hops away (pages linked by randomly selected hyperlinks in the summary section) from the question page as negative samples.

C.2 Demographic Diversity Results

We added a "Diversity" widget that determines the entities[11] (e.g., George Orwell) that capture the nationalities[12] (e.g., United Kingdom). We then

[10] https://pageviews.wmcloud.org/topviews/?project=en.wikipedia.org&platform=all-access&date=last-month&excludes=

[11] https://cloud.google.com/natural-language/docs/analyzing-entities

[12] https://www.wikidata.org/wiki/Wikidata:REST_API

provide suggestions to the authors to include entities from underrepresented countries. However, the questions' demographic diversity distribution did not conform to the population distribution (Equation 2.2), and the entities in the questions showed few country representations.

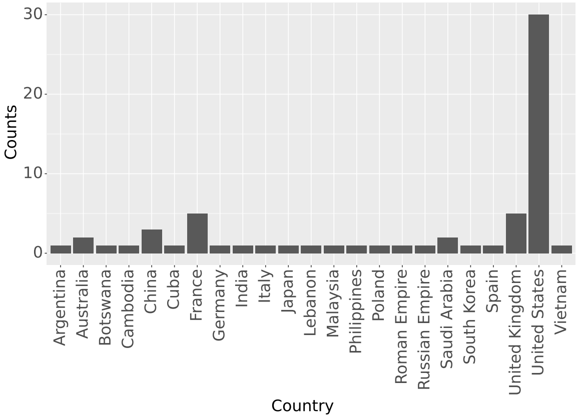


Figure 7: The demographic diversity distribution had negative result as the questions did not contain much nationalities and thus did not conform to population distribution.

C.3 Interface Leaderboard

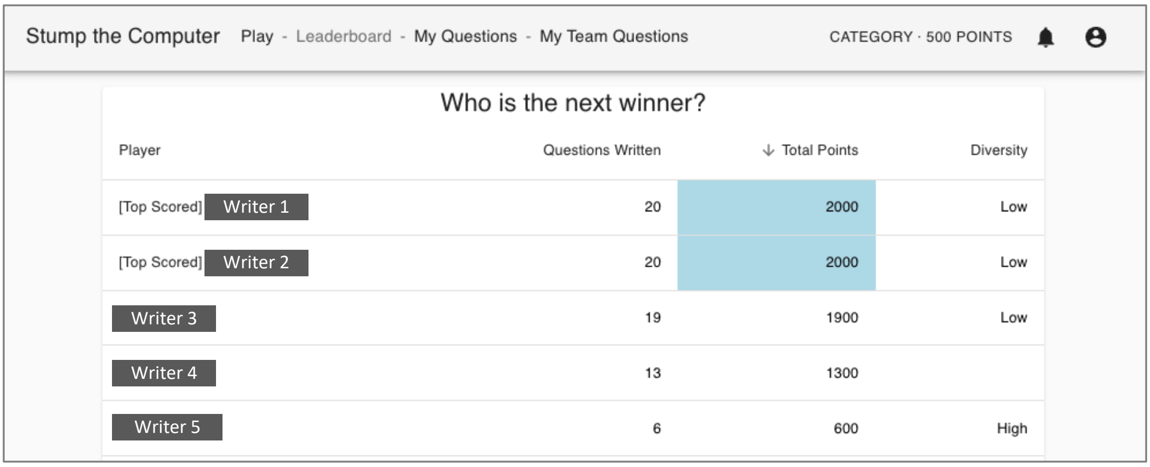
We also build a leaderboard page for writers to keep track of their scores and their diversity score. Figure 8 shows an example of the leaderboard where it displays each writer's name, score, and diversity score.

Question	**Answer**
Art	Questions about works: Mona Lisa, Raft of the Medussa, B) Questions about forms: color, contour, texture, C) Questions about artists: Picasso, Monet, Leonardo da Vinci, D) Questions about context: Renaissance, post-modernism, expressionism, surrealism
Literature Movement	A) Questions about works: novels (1984), plays (The Lion and the Jewel), poems (Rubaiyat), criticism (Poetics), B) Questions about major characters or events in literature: The Death of Anna Karenina, Noboru Wataya, the Marriage of Hippolyta and Theseus
Literary Movement	A) Cross-cutting questions (appearances of Overcoats in novels), B) Common link questions (the literary output of a country/region)
Geography	A) Questions about location: names of capital, state, river, B) Questions about the place: temperature, wind flow, humidity
History	A) When: When did the First World war start?, B) Who: Who is called Napoleon of Iran?, C) Where: Where was the first Summer Olympics held?, D) Which: Which is the oldest civilization in the world?
Science	Questions about terminology: The concept of gravity was discovered by which famous physicist?, Questions about the experiment, Questions about theory: The social action theory believes that individuals are influenced by this theory.
TV and Film	Quotes: What are the dying words of Charles Foster Kane in Citizen Kane?, Title: What 1927 musical was the first "talkie"?, Plot: In The Matrix, does Neo take the blue pill or the red pill?
Music	Singer: What singer has had a Billboard No. 1 hit in each of the last four decades?, Band: Before Bleachers and fun., Jack Antonoff fronted what band?, Title: What was Madonna's first top 10 hit?
Lifestyle	Clothes: What clothing company, founded by a tennis player, has an alligator logo?, Decoration: What was the first perfume sold by Coco Chanel?
Sports	Known facts: What sport is best known as the 'king of sports'? Nationality: What's the national sport of Canada? Sport player: The classic 1980 movie called Raging Bull is about which real-life boxer? Country: What country has competed the most times in the Summer Olympics yet hasn't won any kind of medal?

Table 7: Categories of questions along with the subcategories and corresponding examples.

Question Type	Description	Examples
Lacks Factuality	Requires information is factual	“Trump, the first woman president of the United States, is charged against federal laws” is non factual as the gender of Trump is male
Lacks Specificity (False Presupposition)	Requires more information to be answered with clarity	’What is the color of Flamingo’s feathers?’ is ambiguous as Pink and White could be two possible answers depending on when they are born
Subjectivity	Contains clues that are highly subjective	“What’s the name of Christopher Columbus’s most famous ship?” Possible answers could be either Santa Maria, La Nina, Santa Clara. Also, as “Most famous” can mean many different things, the revised question could be “Which of Columbus’s ships was stripped of its timbers to build a fort called La Navidad in northern Haiti?”
Ambiguity & Multiple acceptable answers	Can be answered with multiple answers	Nikolas Alexandrovitch Romanov, Nikolas II, Nikolai II Alexandrovich Romanov: all of these are acceptable as answers.

Table 8: We list the problematic question types that we ask to annotate. The four types are illustrated with descriptions and examples to help them better understand each question, and help determine whether each question has good quality.



Who is the next winner?

Player	Questions Written	↓ Total Points	Diversity
[Top Scored] Writer 1	20	2000	Low
[Top Scored] Writer 2	20	2000	Low
Writer 3	19	1900	Low
Writer 4	13	1300	
Writer 5	6	600	High

Figure 8: Writer Leaderboard in Interface

Question Type	**Adversarial Type**
Composing seen clues	Contains clues that need to be integrated for the question to be answered
Logic and Calculation	Requires mathematical or logical operators
Multi-Step Reasoning	Requires multiple reasoning steps between entities. For eg: “A building dedicated to this man was the site of the “I Have A Dream” speech.” A reasoning step is required to infer : “I have a dream” speech -> Lincoln Memorial -> Abraham Lincoln
Negation	Contains “not” or “non-” and “no” or any negation entities that may confuse the model to answer
Temporal Misalignment	Contains a specific year, month, or timely event that the model got confused about or does not know.
Location Misalignment	Contains a location that the model got confused about or does now know.
Commonsense Knowledge	Requires information that cannot be answered without commonsense
Domain Expert Knowledge	Requires information that cannot be answered without domain expert knowledge
Novel Clues	Contains information that exists in the question but is not required to answer. These confuse the models.
Crosslingual	Contains multilingual aspects that confuse the model.

Table 9: We list adversarial types (techniques) to determine how each question is using them to stump the models. The annotators are given the description and examples to better understand the reasons why the models may have been stumped. They are expected to tag the examples with the model prediction and question.

Question	**Answer**	**Adversarial Type**	**Question Type**	**Grounding**
What is a fourth of the 5th Bell number, often seen as an unlucky number?	13/Thirteen	Logic & Calculation	Subjectivity	“Unlucky” is a subjective term.
What is the famous meme to come from The Last Dance?	and I took that personally	Commonsense Knowledge, Composing Seen Clues	Multiple Acceptable Answers	The meme can be referred to *many* titles: “Jordan’s Cigar”, “Jordan’s Meme”, ”Laughing Jordan”, and “Crying Jordan”
What substance can cause burns in its gaseous form, lead to vomiting and sweating in high doses, and is the main component by weight in acid rain?	Water	Logic & Calculation, Composing Seen Clues	Specificity	*Many substances* could cause these effects in the novel portion.
Name the title character of the 2024 Best Picture nominee about a fictional conductor who Leonard Bernstein mentored.	Lydia Tar	Temporal Misalignment, Composing Seen Clues	Factuality	2024 Best Picture Nominee *cannot be factually identified* yet
The easternmost state in the U.S. has more than triple its population in lakes and it is known to have good salmon, which state is it?	Alaska	Multihop Reasoning&Location Misalignment	Subjectivity, Specificity	*Good salmon* is subjective, and *easternmost is misleading and it requires relative position* of the author, hence non-specific.

Table 10: We annotated whether each question falls into which adversarial and question type. While being adversarial; some questions lack specificity and factuality. Other questions contained subjectivity and specificity.

Question	**Answer**	**Difficulty Rank**
What is the name of the language which only has 45,900 speakers, allows for word-initial double consonants, and is the official language of an island country with the world’s second-largest regional shark sanctuary?	Chuukese	1
What substance can cause burns in its gaseous form, lead to vomiting and sweating in high doses, and is the main component by weight in acid rain?	Water	2
A large portion of the sequence for the reference genome for the International Human Genome Sequencing Consortium in the human genome project came from a man from which US city?	Buffalo, New York	3
⋮	⋮	⋮
What year is the closest palindromic year to 2001?	2022	88
Which political party governs the country directly south of Botswana?	African National Congress	89

Table 11: Question examples sorted by difficulty value from the IRT model

Question	Answer	γ
What city in England is home to the studios where the British game show that resembles quiz bowl is filmed and where the test of whether an AI is intelligent was proposed?	Manchester	0.193
The 1973 Thomas Rockwell novel for children, about a school boy who loses a gross dare, was written in America. However, a similarly-titled religious edict published in 1521 hails from this old European city of about 80,000.	Worms, Germany	0.192
⋮	⋮	⋮
Who was able to turn men into stone sculptures by just taking a glance at them?	Medusa	0.003
Which political party governs the country directly south of Botswana?	African National Congress	0.003

Table 12: Questions sorted by discriminability(γ) value of the IRT model.

Question	Answer	TFIDF	DPR	CHATGPT
Which political party governs the country directly south of Botswana?	African National Congress	A liberation movement known for its opposition to apartheid, it has governed the country since 1994, when the first post-apartheid election installed Nelson Mandela as President of South Africa.	*Rank 3* Partly due to its Tripartite Alliance with the SACP and the Congress of South African Trade Unions, it has retained a comfortable electoral majority at the national level and in most provinces, and has provided each of South Africa's five presidents since 1994.	The African National Congress (ANC) governs South Africa, which is directly south of Botswana. The ANC has been the ruling party in South Africa since the end of apartheid in 1994.
What company launched their first advertisement campaign during the 1984 Super Bowl that was inspired by George Orwell's dystopian novel "1984"?	Apple	From the introduction of the Macintosh in 1984, with the 1984 Super Bowl advertisement to the more modern Get a Mac adverts, Apple has been recognized for its efforts towards effective advertising and marketing for its products.	***Incorrect Answer: Macintosh*** The first Mac was released in 1984, and was advertised with the highly-acclaimed "1984" ad.	***Correct Answer: Apple*** The launch of Apple's '1984' ad during the Super Bowl is a well-known event in the advertising industry and is often cited as one of the most iconic ads of all time.
A play by Arthur Miller accounts what series of legal proceedings that accused over 200 individuals of consorting with the Devil in late 17th century New England?	Salem Witch Trials	Speakers at the ceremony in August included playwright Arthur Miller and Nobel Laureate Elie Wiesel.	*Rank 2* One other man, Giles Corey, was pressed to death after refusing to enter a plea, and at least five people died in jail.Arrests were made in numerous towns beyond Salem and Salem Village (known today as Danvers), notably Andover and Topsfield.	The question mentions a play by Arthur Miller, which is most likely 'The Crucible' - a dramatized retelling of the Salem Witch Trials. The question also mentions legal proceedings and accusations of consorting with the Devil, which are key elements of the Salem Witch Trials.
What is the title of the 2003 best-selling book that delves into the lives of two young boys in Afghanistan, exploring themes of friendship, betrayal, and redemption against the backdrop of political turmoil?	The Kite Runner	Because its themes of friendship, betrayal, guilt, redemption and the uneasy love between fathers and sons are universal, and not specifically Afghan, the book has been able to reach across cultural, racial, religious and gender gaps to resonate with readers of varying backgrounds.	*Rank 1* Hassan is a successful "kite runner" for Amir; he knows where the kite will land without watching it.	I predicted this answer because I recognized the themes of friendship, betrayal, and redemption as being central to The Kite Runner, which is a well-known and highly acclaimed novel set in Afghanistan. Additionally, the book was published in 2003 and was a best-seller, which further supports this answer.
This disney star had a show named after the state whose capital is Helena. This singer's most recent hit was a song about the brother of Chris Hemsworth. Liam Hemsworth, who is famously known for playing Josh Taylor, married which Disney star?	Miley Cyrus	While filming The Last Song, Cyrus began an on-again, off-again relationship with her co-star Liam Hemsworth in 2009.	*Rank 1* Cyrus, a daughter of country music singer Billy Ray Cyrus, emerged as a teen idol while portraying the title character of the Disney Channel television series Hannah Montana (2006–2011).	The Disney star with a show named after the state whose capital is Helena is Miley Cyrus, who starred in 'Hannah Montana'. Her most recent hit is 'Midnight Sky', which references Liam Hemsworth in the lyrics. Liam Hemsworth was previously married to Miley Cyrus. I predicted this answer because Miley Cyrus is a well-known Disney star who had a show named after a state capital, and she has been in the news recently for her music and personal life, including her relationship with Liam Hemsworth.

Table 13: Example of Comparison between explanations that TFIDF, DPR, and CHATGPT provide for their respective answers. Some evidence shows similar content when predicting the correct answer. Some of the examples have one common clue that hints the models to their predictions, while for other examples, each model has their own groundings for their answers.