Do great minds think alike? Investigating Human-AI Complementarity for Question Answering

Maharshi Gor[1] **Tianyi Zhou**[1] **Hal Daumé III**[1,2] **Jordan Boyd-Graber**[1]

[1]University of Maryland [2]Microsoft Research

Abstract

The emergence and steady progress of large language models (LLMs) has caused credulous commentators to claim AIs had obtained superhuman ability on natural language processing (NLP) tasks such as textual understanding and reasoning. This work investigates this question and explores the complementarity of skills between humans and AI in answering complex entity-seeking questions of diverse topics and types. This work proposes a framework that builds on top of multidimensional item response theory (MIRT), a tool from the psychological literature. MIRT learns representations of both the difficulty of questions and the skills of agents—both human and computer—to predict their performance on each question. Humans can answer questions with fewer clues and higher accuracy than most QA systems except GPT-3, where humans only do better with more abstract clues that require higher-order multi-hop reasoning. This highlights both where future QA work can best focus and how to build collaborative human-computer QA systems.

1 Introduction

The natural language processing (NLP) community has long focused on developing systems capable of *emulating* human behavior, treating human performance as a ceiling for NLP models. The latest wave of LLMs has turned the discussion to supremacy: models are purportedly acing tests (OpenAI, 2023; Liu et al., 2023) that many humans find challenging.[1] And there are indeed areas where computers seem to have human-level ability.

For NLP, the most notable example of this was IBM Watson's *tour de force* performance Ferrucci et al. (2010) on *Jeopardy!*. While Watson defeated the two human specimens on the stage, to the best of our knowledge there has not been a thorough, quantitative examination of the relative strengths and weaknesses of human vs. computer on question answering, particularly with the new panoply of LLMs available in 2023. This paper seeks to close that gap with an examination of human and computer ability on question answering.

We choose question answering (QA) because it is a complex natural language processing task requiring contextual question understanding, retrieval of relevant information, and synthesizing that information to answer a question. This is in stark contrast to problems such as sentiment analysis where both humans and modern AI systems are independently capable of saturating performance metrics on common benchmarks; moreover, answering tricky questions has captured the human fascination. Like Watson, we use the trivia domain: questions are carefully crafted to probe the knowledge and reasoning ability of human players. Unlike Watson, rather than comparing one AI against two human/teams on a couple dozen questions, we compare ~ 30 AI systems against 20 humans on thousands of questions.

Further elaborated in § 2.1, we use a QA format (He et al., 2016; Rodriguez et al., 2019) specifically designed to enable effective comparison between QA agents. Like Watson, these are trivia questions, but the questions we use are designed to challenge even the strongest trivia whizzes (in contrast, Watson's virtuoso human opponents could easily answer the vast majority of the questions on "normal" *Jeopardy!*).

To analyze the questions and answers, we build on item response theory (IRT,§2.2), which was first introduced in the field of Psychometrics (Santor and Ramsay, 1998). A classical IRT model, which assesses agent skills and question characteristics (difficulty, discriminability) based on binary response correctness, typically relies on a one-dimensional representation. This approach lacks

[1]As should hopefully be clear from the rest of the paper, we are highly dubious of these claims, particularly on multi-choice tests with copious study material online. But this is outside the main scope of *this* paper.

the capacity to effectively represent the multimodal nature of response distributions. Additionally, its naïve multidimensional extension suffers from non-identifiability, where different combinations of difficulty and skills can yield identical responses. Furthermore, IRT's reliance on question identifiers for representation limits its applicability for assessing difficulty of unseen questions.

To overcome these limitations, we propose a new framework: Content-aware, Identifiable, and Multidimensional Item Response Analysis (CAIMIRA, pronounced as ***Chimera***) in Section 3. We apply CAIMIRA to responses collected from trivia players and a wide range of QA models (§ 4) over our questions, and provide a thorough analysis of question and agent characteristics (§ 5).

Our analysis reveals five latent axes that models characteristics of questions and agents in consideration: *Geography*, *Culture/History*, *Scientific Reasoning*, *Periphrastic Narrations*, and *Entity Specificity*. There is a notable difference between human and ai skill-set. Overall, humans are more consistent across the axes than most QAmodels. All QA systems, including GPT-4, struggle with the questions that some humans ace at: one requiring scientific reasoning and/or character descriptions with indirect speech, or *periphrasis*. In contrast, very-large scaled models make fewer errors than humans for questions with high *entity specificity*, which involves *knowing* historical or geographical facts; more so, they also match the retrievers' recall scores just using their parametric memory. We also find that questions strongly associated with static facts (e.g. Geography), are generally easier, while questions requiring scientific reasoning or indirect speech have a higher degree of variance in question difficulty, suggesting that the benchmarks involving latter can better discriminate more effective agents from the worse ones.

2 Background and Preliminaries

This section describes our dataset, the source of human question answering data, and preliminaries of our methodology. These data are the input into our CAIMIRA framework, which we develop in Section 3.

2.1 QUIZBOWL: Where Trivia Nerds Practice

There are many QA datasets to choose from; Rogers et al. (2023) enumerates a plethora in their thorough review, delineating between "information seeking" and "probing" questions (Min et al., 2020; Yu et al., 2022). Our overarching goal is to identify similarities and differences between system and human answers; hence, we focus on probing-style questions as they are more diverse, less prone to ambiguity or false presuppositions, and are designed to be particularly challenging (at least for humans). More importantly, we need questions with many examples of human answers. While humans do not sit around answering Google queries (Kwiatkowski et al., 2019) for fun, they do answer trivia questions for both for the intrinsic fun and to prepare for future trivia competitions.

We draw on the dataset from He et al. (2016), which is popularly known as "Protobowl". This dataset is based on the Quizbowl QA setting (Boyd-Graber et al., 2012, QB). To our knowledge, it is the only open source QA dataset that contains records of many human players of varying levels of expertise answering questions across different categories.

Quizbowl (Rodriguez et al., 2019), the source of questions for ProtoBowl, is a trivia game where players can **interrupt** questions to answer and answering earlier is better. This can better discriminate players' skills because the clues progress from hard to easy, culminating with a "giveaway" hint at the end of the question. Often, the sequence of clues reveals more information or helps disambiguate possible references and interpretations at each step. In contrast to "all or nothing" QA, incremental QB questions—progressively exposing an agent with a series of clues of increasingly enriched information—helps pinpoint the clues necessary for user a to answer question q.

We collect multiplayer logs from questions played across all categories. The best players have deep knowledge and excellent lateral thinking skills (Jennings, 2006). Player logs record question metadata, including question category (e.g. History) and target player level (e.g., college novice), time taken to answer the question, answer string, and the correctness ruling by the "Protobowl" platform.

2.2 A review of Item Response Theory (IRT)

We compare human users and AI systems in Quizbowl QA setting by investigating their skills and complimentarity on varied questions using Item Response Theory (IRT), a framework typically used to analyze human responses (ruled as correct or incorrect) to a set of questions (or, "items"). It

is widely adopted in psychometrics (Morizot et al., 2009), medical education (Downing, 2003), and other fields for developing tests for human subjects.

In the context of this work, IRT assumes question-answer pairs that form an evaluation set, subjects spanning humans to QA systems, and responses rulings of these agents. The objective of a simple IRT system is to jointly model agent skills and question characteristics that best predicts the responses rulings (Baker and Kim, 2004).

We first review the simplest formulation of an IRT model which uses a scalar to represent agent skill (s_i) and question difficulty (d_j):

$$p(U_{i,j} = 1 \mid s_i, d_j) = \sigma(s_i - d_j), \qquad (1)$$

where $\sigma(z) \triangleq 1/1+e^{-z}$ is the logistic function and $U_{i,j} \in \{0, 1\}$ is the binary ruling of the i^{th} agent's response to the j^{th} question.

Existing work in NLP using IRT mainly relies on simple uni-dimensional models (Lalor et al., 2019) to represent question characteristics, which is adequate in certain contexts. The model implicitly assumes a monotonicity of the parameters: a History question q_h with a higher difficulty than a Science question q_s ($d_h > d_s$) would entail that agents who get q_s right must also get q_h right, no matter their expertise. While this simplified assumption is useful in some cases, it cannot capture the *diversity* of questions and agent capabilities and their matching.

Multidimensional Latent IRT (MIRT). To relax the monotonicity assumption, and model multi-factor characteristics, contemporary work (Chalmers, 2012) proposes a multidimensional *discriminability* vector $\boldsymbol{\alpha_j}$ of item-j to interact with a multidimensional agent *skill* $\mathbf{s_i}$. The resulting MIRT model has two question characteristics, i.e., a scalar *difficulty* d_j and an m-dimensional $\boldsymbol{\alpha_j}$. The objective is then computed as:

$$p(U_{i,j} = 1 \mid \mathbf{s_i}, d_j, \boldsymbol{\alpha_j}) = \sigma(\mathbf{s_i}^\intercal \boldsymbol{\alpha_j} - d_j). \qquad (2)$$

The discriminability $\boldsymbol{\alpha_j}$ aims to capture how sensitively the correctness probability changes with each dimension of the agent skill $\mathbf{s_i}$. Although $\boldsymbol{\alpha_j}$ can be used to match the agent expertise on different dimensions, the difficulty d_j is dimension agnostic. Moreover, there are limited constraints on the values in $\boldsymbol{\alpha_j}$, allowing multiple or even infinite different but observationally equivalent choices of $\mathbf{s_i}$, d_j, $\boldsymbol{\alpha_j}$, making them *non-identifiable*. Lastly, the model's reliance solely on unique identifiers, instead of textual *content*, hampers its ability to estimate latent characteristics of new questions, necessitating frequent model retraining—an unscalable approach. This limitation also disallows the model to capture the effect of linguistic nuances of the questions on its characteristics.

We address these shortcomings in our proposed framework, CAIMIRA (§ 3), by extending both the skill s_i and difficulty d_j to be multidimensional and incorporating enhancements to improve model generalizability, interpretability and address the non-identifiability.

3 Bootstrapping Item Response Theory with CAIMIRA

This section describes our proposed approach—Content-aware, Identifiable, and Multidimensional Item Response Analysis (CAIMIRA)—that addresses the limitations of MIRT (§ 2.2) by making three primary changes. In particular, we (i) replace *discriminability* $\boldsymbol{\alpha_j}$ with a normalized weight-vector *relevance* $\mathbf{r_j}$ that aggregates the differences between *skill* and *difficulty* $(\mathbf{s_i} - \mathbf{d_j})$ over the dimensions that are most relevant to the question, to a scalar $((\mathbf{s_i} - \mathbf{d_j})^\intercal \mathbf{r_j})$, (ii) mean-shift *difficulty* to zero, which resolves the identifiability issue between difficulty and skills, and (iii) learn a transformation from the question's dense representations to its *relevance* and *difficulty* vectors, achieving content-awareness. We decompose the information captured jointly in MIRT's item characteristics *discriminability* $\boldsymbol{\alpha_j}$ and a scalar difficulty d_j (which performs as an intercept in Eq. (2)) into more controlled multidimensional characteristics *relevance* $\mathbf{r_j}$ and *difficulty* $\mathbf{d_j}$ in CAIMIRA. The CAIMIRA objective is:

$$p(U_{i,j} = 1 \mid \mathbf{s_i}, \mathbf{r_j}, \mathbf{d_j}) = \sigma\left[(\mathbf{s_i} - \mathbf{d_j})^\intercal \mathbf{r_j}\right] \qquad (3)$$

where $\mathbf{s_i}, \mathbf{r_j}, \mathbf{d_j} \in \mathbb{R}^m$ are agent skills, and question characteristics respectively.

3.1 Introducing question *relevance* $\mathbf{r_j}$

relevance $\mathbf{r_j}$ is a measure of each latent-factor's contribution to the overall answerability of the j^{th} question. It aggregates the m-dimensional latent *scores*—differences between $\mathbf{s_i}$ and $\mathbf{d_j}$ to a scalar value $(\mathbf{s_i} - \mathbf{d_j})^\intercal \mathbf{r_j}$, that is used to compute $p(U_{i,j} = 1)$. As part of regularization, it is structured as a probability distribution across m latent

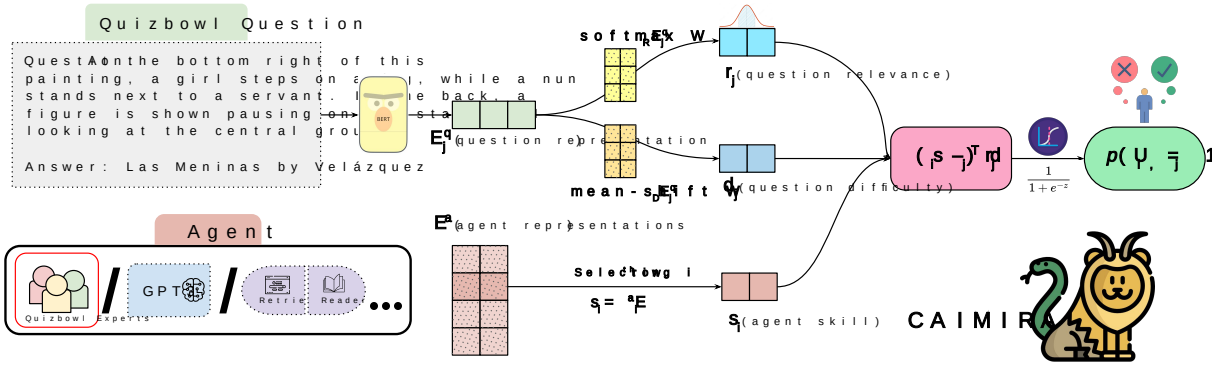

Figure 1: A pipeline of CAIMIRA. It predicts the probability of agent-i correctly answering question-j using a model in Eq. (3), where the question's relevance $\mathbf{r_j}$ and difficulty $\mathbf{d_j}$ are multidimensional and computed by linear transformations of the question embedding $\mathbf{E}_j^q$ (§ 3.3), and the agent skill $\mathbf{s_i}$ is extracted from a learnable agent embedding matrix $\mathbf{E}^a$. Unlike MIRT, $\mathbf{r_j}$ is a probability distribution computed from the raw reference $\mathbf{r'_j}$ and improves the interpretability of the multidimensional model (§ 3.1); $\mathbf{d_j}$ is achieved by zero centering of the raw difficulty $\mathbf{d_j}$, which addresses the identifiability issue of $\mathbf{s_i}$ and $\mathbf{d_j}$ in $\mathbf{s_i} - \mathbf{d_j}$ (§ 3.2).

factors, ensuring that the sum of the values equals ($\sum_{k=1}^{m} \mathbf{r_{jk}} = 1$), and each value is nonnegative. This is achieved by applying softmax function to the **raw relevance** $\mathbf{r'_j}$ (Figure 1).

The probability-simplex constraint to $\mathbf{r_j}$ improves the interpretability of the latent scores $(\mathbf{s_i} - \mathbf{d_j})$, and prevents confounding when comparing agent-skills and question-difficulties. For instance, in a two-dimensional latent space, consider a moderately difficult history question. Two agents with identical history skills but differing scientific reasoning skills should have an equal likelihood of answering correctly if the question predominantly tests history knowledge. This is realized by assigning a high relevance to the history dimension (near 1) and a low relevance to the scientific reasoning dimension (near 0).

3.2 Zero Centering of *difficulty* $\mathbf{d_j}$

The difference between the agent skill and the question difficulty $(\mathbf{s_i} - \mathbf{d_j})$ determines the correctness probability, not just their numerical values, allowing different optimal choices of skill $\mathbf{s_i}$ and difficulty $\mathbf{d_j}$ that produce the same probability, i.e., $\mathbf{s_i}$ and $\mathbf{d_j}$ are non-identifiable. To alleviate this issue, we regulate the **raw difficulty** $\mathbf{d'_j}$ of each question q_j to have a zero mean over each dimension without affecting the correctness probability, i.e., $\mathbf{d_j} = \mathbf{d'_j} - 1/n_q \sum_{j=1}^{n_q} d'_j$, where n_q is the total number of questions. This limits the range of skills and difficulty, and also allows us to compare the agent skills and question difficulties across the latent dimensions, which was harder in MIRT and required additional post-processing steps.

3.3 From MIRT to Content-Aware CAIMIRA

The CAIMIRA framework extends MIRT by making use of the actual question texts (content-aware) to compute their characteristics and handle new questions at inference (cold-start friendly). It learns linear transforms from the its embedding vector $\mathbf{E}_j^q$ to its raw characteristics $\mathbf{r'_j}$ and $\mathbf{d'_j}$, which are then normalized to obtain the final characteristics $\mathbf{r_j}$ and $\mathbf{d_j}$. Mathematically,

$$\mathbf{r'_j} = \mathbf{W}_R\, \mathbf{E}_j^q + \mathbf{b}_R, \quad \mathbf{d'_j} = \mathbf{W}_D\, \mathbf{E}_j^q, \tag{4}$$

$$\mathbf{r_j} = \text{softmax}(\mathbf{r'_j}), \qquad \mathbf{d_j} = \mathbf{d'_j} - \frac{1}{n_q} \sum_{j=1}^{n_q} \mathbf{d'_j}, \tag{5}$$

where $\mathbf{W}_R, \mathbf{W}_D \in \mathbb{R}^{m \times n}$ and $\mathbf{b}_R \in \mathbb{R}^m$. They and the embedding matrix $\mathbf{E}^a$ of agent skills ($\mathbf{s_i} = \mathbf{E}_i^a$) are the parameters we train for CAIMIRA.

The embedding $\mathbf{E}_j^q$ is a high-dimensional representation of the question, which can be obtained on-fly using a pre-trained transformer encoder like BERT, or a sparse tf-idf representation.

4 Experimental Setup

In this section, we describe the process for gathering the responses from human agents and QB systems for our questions, and determining their rulings. We also describe our process to analyze the latent characteristics learnt by CAIMIRA through these responses.

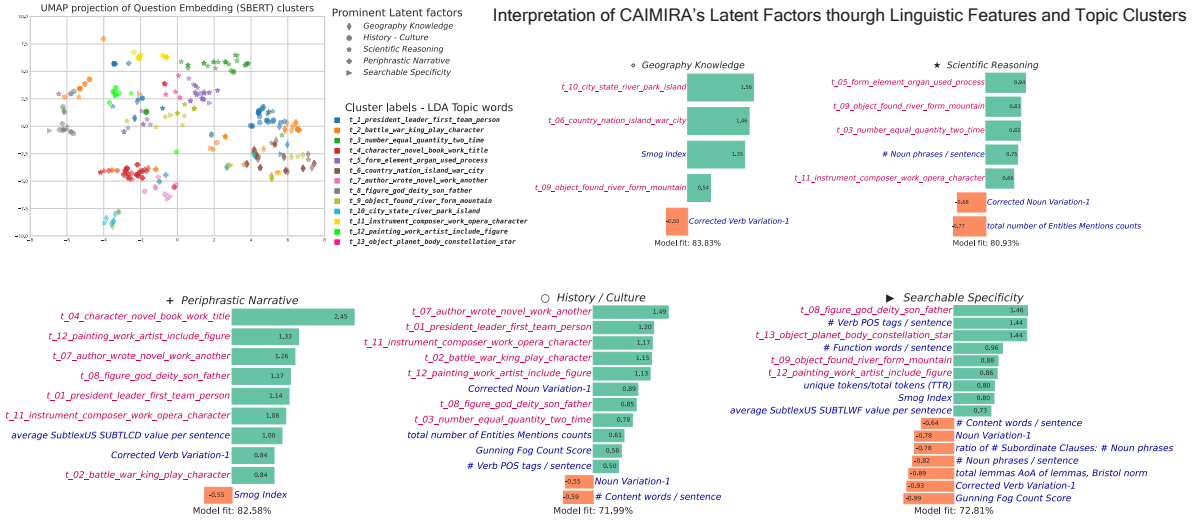

Figure 2: To understand what interpretable features that contribute to latent factors, we turn to Logistic Regression. We construct a one-vs-all binary classification task for each latent factor. A question is labelled positive for k^{th} latent factor if the corresponding $\mathbf{r_{jk}} > 0.6$. We report the model fit and the statistically significant features that contribute to it. This shows that predicting the relevance just from the SBERT embeddings is effective.

Data Source. Questions in protobowl are multi-sentence *clues* about a certain entity or a concept (the answer) that the player need to respond with. Protobowl logs maintains this clue-level information for each question. We consider these as different entries in our datasets to learn if providing *more* clues has a similar effect on machine responses as that of trivia experts. Typically, a question on average has 4 clues. Consider a question `q31` with 3 clues. We maintain three entries in our dataset for this question, one corresponding to the question text till the end of each clue: [`q31_1`, `q31_2`, `q31_3`]. Player responses to `q31` after the second clue are recorded under `q31_2`. To align with models prompted with all clues, we backfill `q31_1` responses with the `q31_2` response if `q31_2` is incorrect, and subsequent unseen clues are backfilled with `q31_1` responses if `q31_1` is correct. In total, we gather 3042 entries in our dataset.

Human players. Raw human responses are sparse: many players only answer a few dozen questions, not having a strong overlap, making it hard to contrast human–AI complementarity. We prune the logs and retaining questions that have at least 100 human responses and backfill the entries. We consider 5 certain anonymized human players who have answered at least 1500 questions each. We also consider *groups* of multiple human players as agents. A grouped human agent behaves like a single agent, but is composed of multiple human players. The final response of a grouped human agent is determined by the majority response of the players in the group if more than one player attempts that question. If multiple players attempt the question and there is no majority, we randomly select one of the responses.[2] The grouping also helps reduces the sparsity of human responses keeping higher amount of overlap in the set of attempted questions among human agents. This also enables us to measure the effect of group size on "agent" skill. To determine the ruling of a grouped-human agent, we take the majority response as their final guess and evaluate it against the answer. In this study we consider group sizes of 1, 5, 10 and 15, and sample 5 of each, making a total of 20 human agents.

4.1 AI agents

We choose a diverse set of QA systems that span a range of complexity and training sets to capture skill differentials between models. Choosing varied systems not only mimics the novice-expert divergence in humans but also gives us insight into the individual strengths of particular training and modeling paradigms. We briefly introduce the model-based agents below that we use to generate responses for our dataset.

[2]This is a naive group setting. We leave the exploration of more sophisticated real-like group settings to future work.

Retrievers as QA agents. Our retrievers, which index Wikipedia documents, respond with the top k documents (where k = 1, 3, 5, 10) most relevant to the question. We employ two types: dense and sparse. The dense retriever, CONTRIEVER (Izacard et al., 2021), is pretrained via unsupervised contrastive learning on a mix of Wikipedia and CC-Net data, then fine-tuned on MS-MARCO (Campos et al., 2016). The sparse retriever utilizes the BM25 algorithm (Robertson and Zaragoza, 2009) and Anserini's implementation and index (Lin et al., 2021). We also test a title-retriever, assuming the document title as the query answer. Retrievers are evaluated on recall-based accuracy, with a point scored if the answer appears within the top k documents for context-retrievers, or in the title of the top k documents for the title-retriever.

Large Language Models (LLMs). We evaluate an array of LLMs, grouped below by their training / scale. All models are evaluated in a zero-shot manner (no finetuning over QB questions).
Base Models: The models are exclusively trained on an unsupervised CausalLM objective: OPT (Zhang et al., 2022), GPT-Neo (Black et al., 2021) and Pythia (Biderman et al., 2023)
Benchmark Instruction Tuned (IT) Models: LLMs fine-tuned on tasks with natural instructions over each benchmark; OPT-IML (Iyer et al., 2022), T0, T0pp (Sanh et al., 2021), Flan-T5 (Chung et al., 2022) and Flan-UL2 (Tay et al., 2022).
Very Large-Scaled Models: Llama-2 (70 billion parameters) (Touvron et al., 2023) and Falcon (40 billion parameters) (Almazrouei et al., 2023) and its instruction tuned variant. Due to limited information on their training data mixtures, direct comparisons with other models are challenging. Nevertheless, we include these large-scale models to gauge their performance relative to humans.
Closed-Sourced Model-Based APIs: OpenAI's ChatGPT (Ouyang et al., 2022) and GPT-4 Turbo (OpenAI, 2023)

None of the Transformer-based models, including those pretrained on QA datasets like TriviaQA, are specifically finetuned on QB; we adhere to the standard in-context learning practice (Brown et al., 2020),providing a task instruction followed by concatenated QA pair demonstrations.

Retriever-augmented Generative Models. Following the RAG paradigm from (Lewis et al., 2020) for open-domain QA, we first retrieve Wikipedia documents relevant to the questions, then employ a generator model for short answer generation. Our retrievers include dense CONTRIEVER and a sparse passage retriever (BM25). For the retriever, we use both a dense retriever (CONTRIEVER) as well as a sparse passage retriever that uses BM25 to encode documents. The generator models used are FlanT5-XL (Chung et al., 2022), constrained by a 512-token context limit and utilizing the top 3 documents, and Flan-UL2 (Tay et al., 2022), an instruction-tuned UL2 with a 2048-token receptive field, capable of handling all 10 documents.

Answer Equivalence. Traditional exact match metric often misses alternative answer forms that are actually correct (Bulian et al., 2022). To better handle this, we adopt a fuzzy match evaluation using multiple answer aliases (Si et al., 2021) where we set a threshold on the character-level match rate between the predicted answer and gold answer and judge all predictions above the threshold as correct. The threshold is tuned against human judgments on a small development set.

4.2 CAIMIRA Setup

We train a 5-dimensional CAIMIRA model to learn the latent characteristics of questions and agents. To get question embeddings $\mathbf{E}_j^q$, we use SBERT (Reimers and Gurevych, 2019) fine-tuned from MPNET (Song et al., 2020). We also include the answer and the first paragraph of its related Wikipedia page in the text input to SBERT. This approach helps in capturing a more comprehensive context of the question. The trainable parameters are fit using mini-batch stochastic gradient descent to minimize the cross entropy loss between the predicted likelihood $p(U_{i,j})$ and the true ruling of the response $U_{i,j}$ as in Equation 3. We use Adam optimizer (Kingma and Ba, 2014) without weight decay, and with a learning rate of 0.005.

Interpretation of Latent Factors We employ Logistic Regression to establish an interpretable link between question text and its characteristics learnt by CAIMIRA, using categorical topic cluster labels, clue counts, and extensive linguistic features (as per (Lee et al., 2021)) as inputs. This approach outputs a probabilistic *relevance* measure, effectively connecting complex linguistic aspects with practical question *relevance* assessment. To interpret the *relevance* $\mathbf{r_j}$, we adopt the approach from (Gor et al., 2021), performing logistic regression analysis for each latent factor separately, resulting

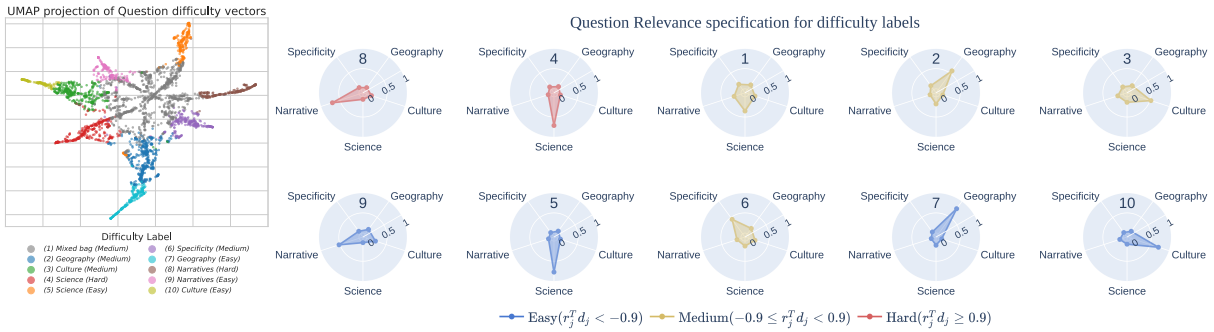

Figure 3: Relevance radar plot for questions clustered in the BERT-embedding space. The color of the plot represents *effective difficulty* $\mathbf{r_j}^\intercal \mathbf{d_j}$ of each cluster.

in binary labels for every question for each dimension.

Our features, derived from a KMeans clustering 768-dimensional SBERT embeddings, result in 13 distinct clusters (Figure 2) for nuanced question categorization. Labeled LDA (Ramage et al., 2009) generates topic-specific words for each cluster, enriching interpretability. The array of linguistic features span advanced semantic, discourse-based, and syntactic elements, providing a rich and multi-faceted representation of the questions. While inspecting dimension k, the output label is 1 if $r_{jk} > 0.6$ and 0 otherwise. Figure 2 lists the most contributing features for each dimension that is statistically significant.

5 Question and Agent Analysis

In this section, we lay out interpretations of the latent dimensions of CAIMIRA using *relevance* in subsection 5.1, and analyze patterns in question difficulties and agent skills in subsection 5.2.

5.1 What are the latent factors?

Figure 2 highlights key attributes of latent factors in our analysis. The first dimension is chiefly characterized by topics in *Geography* (Political or Geological), with a positive correlation to the Smog Index, indicating complexity in text readability due to sentence length and polysyllabic words. Common Geography terms like "attempted", "civilians", and "foreigners" are significant contributors.

The second dimension, *History and Culture*, predominantly involves questions about Authors, Composers, Artists, and Leaders. Notably, there's a higher presence of Entity mentions, reflecting the focus on individuals and their works.

The third axis, *Scientific Reasoning*, centered around scientific phenomena and concepts (e.g., "slope" in mathematics), differs in question style as well. Despite a higher count of noun phrases, the lower number of entity mentions and increased use of multi-sense words poses a challenge to retrieval systems and smaller LLMs. For instance, The question expecting "Matter" as the answer is phrased as *"The density parameter for the non-relativistic form of this falls off with the cube of the scale factor."*

The fourth latent axis, though related to literary works on surface form, is majorly governed by *periphrasis*, or indirect speech, often describing a narrative without direct references to named entities or key phrases. Questions associated with this axis predominantly involve some form of *narration*, typically in a fictional realm: plot of a literary work, events of a music video. This style, involving higher verb variation and contextual diversity but lower Smog Index, is a common source of difficulty in Quizbowl, for both humans and machines. (Rodriguez et al., 2019).

The final dimension, *Entity Specificity*, pertains to questions specifically targeting a particular entity, and also involving well represented and searchable key phrases. Few QB questions score highly on this axis. Retrievers and systems based on them exhibit greater proficiency in this area Figure 4.

5.2 Interpretation of Agent skills

To analyze human and QA model skills across latent dimensions and correlate them with accuracy, we cluster question difficulty vectors using Kmeans and calculate average agent accuracy per cluster. Clusters' *relevance* to the latent dimensions and agent accuracy are depicted in Figure 3 and **??**, respectively.

Out of 10 difficulty clusters, some like *Science* and *periphrasis* contain both challenging (Clusters 8 and 4) and easier (Clusters 9 and 5) subsets, with AI systems, including GPT-4, struggling in the

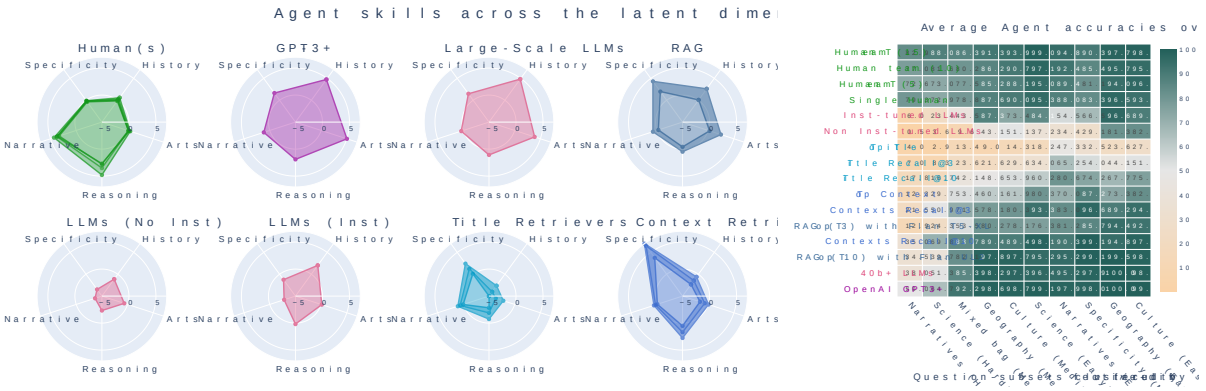

Figure 4: Left (radar plots) shows the average skills of our agents by their categories across our five latent factors. Right (heatmap) shows accuracies of the agents on questions clustered by their difficulty vector.

harder ones.

Geography and *Culture/History* clusters show a mix of easy to medium difficulty. Cluster 1, diverse across dimensions, emerges as the third hardest, where only large-scale models and comprehensive retriever-based systems excel.

In Clusters 2 and 6, humans achieve 87% accuracy, but GPT-4 reaches 98.6%. These clusters, characterized by higher *Specificity* and less *periphrasis*, often involve static factual content, e.g., national capitals.

Interestingly, the Title Retriever (Top 1), despite being the least effective overall, scores 47% in Cluster 9, comprising easier narrative questions. This is attributed to these questions having more clues (4.2 on average) than their tougher counterparts in Cluster 8 (1.7), suggesting retriever systems' ability to utilize additional context for better guessing and resilience against confounders.

6 Related Work

6.1 Adoption of IRT in NLP Evaluation.

Current performance evaluation paradigms in both machine and human QA fail to induce robust, natural partitions in evaluation datasets.

The evaluation is carried out as a sum of parts, treating each question as an independent single transaction rather than considering relative differences between test set items. To remedy this, Lalor et al. (2019) introduce the IRT ranking method from the education testing community as a fruitful new paradigm for NLP evaluation. Rodriguez et al. (2021) argue for the adoption of IRT as the de facto standard for QA benchmarks, demonstrating its utility in guiding annotation effort, detecting annotator error, and revealing natural partitions in evaluation datasets. Together with fine-grained human and machine Protobowl responses, IRT serves as a framework with which to uncover quantitative and qualitative differences between QA agents that would otherwise be washed out due to the unidimensionality of standard QA benchmarks solely implementing mean answer accuracy over test sets.

6.2 Ideal Point Models

Item Response Theory (IRT) and Ideal Point Models (IPM) are two prominent statistical models used in different fields for various purposes.Both models deal with the analysis of preferences or abilities, but their applications and theoretical underpinnings show significant differences. IRM, commonly used in educational testing, evaluates an individual's ability based on their responses to questions, assuming a unidimensional trait (De Ayala, 2013). In contrast, IPM, often applied in political science, assesses individuals' positions on a spectrum, such as political ideology, based on their choices or votes (Clinton et al., 2004). A critical similarity lies in their use of probabilistic frameworks to model responses or choices, allowing for uncertainty in measurements. However, while IRM focuses on the latent trait of an individual, IPM emphasizes the spatial representation of preferences, offering a more multidimensional perspective. This review underscores the need for a nuanced understanding of these models to leverage their strengths in respective domains.

6.3 Human-AI Complementarity and Collaboration

An increasingly popular line of work in NLP studies complementing human capabilities with language models. This often calls for analysis of

human-AI complementarity where we aim to combine the strengths of both humans and AI systems. One common application for such collaboration is creative writing. Lee et al. (2022) recruited human writers to interact with GPT-3 for collaborative writing, where humans can get suggestions from GPT-3 and make further edits. Padmakumar and He (2021) deploy a language model in the loop for modifying user-selected spans to make the draft more descriptive and figurative. Apart from leveraging the "writing" capabilities of language models, in a question-answering context, Feng and Boyd-Graber (2018) recruit experts and novices to play trivia games with AI systems as teammates. Given model predictions, humans needed to decide when to trust the model outputs. He et al. (2022) studied the information retrieval task with human-AI collaboration, where humans are given model-generated queries to help them search through Wikipedia for answers. Our work differs from these papers in two important aspects: firstly, we study how to best combine both human and model predictions by leveraging their complementary skill set rather than letting humans act on the model predictions; secondly, we attempt to model and understand human skills as compared to models, which is largely ignored in previous work, but is crucial for gaining insights into human-AI complementarity.

7 Discussions and Conclusions

Our study, utilizing the proposed CAIMIRA framework, provides insights into how humans and AI systems, complement each other in QA tasks. GPT-4 exhibits impressive proficiency in tackling straightforward questions or complex inquiries that benefit from ample contextual information, closely mirroring human performance. However, when confronted with intricate questions marked by indirect speech and a scarcity of clues, GPT-4 encounters challenges. Similarly, retriever-based systems excel when presented with a wealth of clues but falter in scenarios where indirect speech lacks sufficient contextual detail.

In stark contrast, human participants demonstrate remarkable prowess in deciphering single-clue questions laden with indirect speech, surpassing the capabilities of GPT-4. It highlights the need for datasets that evaluate a model's ability to grasp implicit context, especially as NLP evolves toward conversational agents and real-world problem-solving.

Furthermore, a compelling argument emerges for instructed-tuned and RLHF-based models, which purportedly exhibit human-like behavior. However, they still fall short in invoking the intuitive reasoning that humans possess. Humans have the ability to *seek* additional context, without being instructed, when their initial hypotheses fail, a skill not yet present in AI systems *yet*. This gap emphasizes the importance of further research in this area.

Overall, these findings underscore the intricate interplay between human intuition and AI's analytical capabilities when it comes to comprehending and responding to complex language, highlighting the progress made in AI while acknowledging the unique strengths of human cognition.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

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

Stella Biderman, Hailey Schoelkopf, Quentin G. Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *International Conference on Machine Learning*.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow.

Jordan Boyd-Graber, Brianna Satinoff, He He, and Hal Daume III. 2012. Besting the quiz master: Crowdsourcing incremental classification games.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Jannis Bulian, C. Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *Conference On Empirical Methods In Natural Language Processing*.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, L. Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *COCO@NIPS*.

R Philip Chalmers. 2012. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48:1–29.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review*, 98(2):355–370.

Rafael Jaime De Ayala. 2013. *The theory and practice of item response theory*. Guilford Publications.

Steven M Downing. 2003. Item response theory: applications of modern test theory in medical education. *Medical education*, 37(8):739–745.

Shi Feng and Jordan L. Boyd-Graber. 2018. What can ai do for me?: evaluating machine learning interpretations in cooperative play. *Proceedings of the 24th International Conference on Intelligent User Interfaces*.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3).

Maharshi Gor, Kellie Webster, and Jordan Boyd-Graber. 2021. Toward deconfounding the effect of entity demographics for question answering accuracy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5457–5473, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent modeling in deep reinforcement learning.

Wanrong He, Andrew Mao, and Jordan Boyd-Graber. 2022. Cheater's bowl: Human vs. computer search strategies for open-domain qa. In *Findings of Empirical Methods in Natural Language Processing*.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv: 2212.12017*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*

Ken Jennings. 2006. *Brainiac: adventures in the curious, competitive, compulsive world of trivia buffs*. Villard.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research.

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 Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2019, page 4240. NIH Public Access.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv: 2304.03439*.

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.

Julien Morizot, Andrew T Ainsworth, and Steven P Reise. 2009. Toward modern psychometrics. *Handbook of research methods in personality psychology*, 407.

OpenAI. 2023. Gpt-4 technical report. *PREPRINT*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Vishakh Padmakumar and He He. 2021. Machine-in-the-loop rewriting for creative image captioning. In *NAACL*.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Conference on Empirical Methods in Natural Language Processing*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

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)*.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv: Arxiv-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).

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, S. Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan D. Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng-Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, T. Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *International Conference on Learning Representations*.

Darcy A Santor and James O. Ramsay. 1998. Progress in the technology of measurement: Applications of item response models. *Psychological Assessment*, 10:345–359.

Chenglei Si, Chen Zhao, and Jordan L. Boyd-Graber. 2021. What's in a name? answer equivalence for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier García, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, H. Zheng, Denny Zhou, N. Houlsby, and Donald Metzler. 2022. Ul2: Unifying language learning paradigms. *International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schel-

ten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv: 2307.09288*.

Xinyan Velocity Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Crepe: Open-domain question answering with false presuppositions.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.