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
mgor@cs.umd.edu

Abstract

This study examines question-answering (QA) abilities across human and AI agents. Our framework CAIMIRA addresses limitations in traditional item response theory, by incorporating multidimensional analysis, identifiability, and content awareness, enabling nuanced comparison of QA agents. Analyzing responses from ~ 30 AI systems and 155 humans over thousands of questions, we identify distinct knowledge domains and reasoning skills where these agents demonstrate differential proficiencies. Humans outperform AI systems in scientific reasoning and understanding nuanced language, while large-scale LLMs like GPT-4 and LLAMA-2-70B excel in retrieving specific factual information. The study identifies key areas for future QA tasks and model development, emphasizing the importance of semantic understanding and scientific reasoning in creating more effective and discriminating benchmarks.

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, an early notable example of was IBM Watson's *tour de force* performance Ferrucci et al. (2010) on *Jeopardy!*. While Watson defeated the two humans on stage, to the best of our knowledge, a thorough, quantitative examination of the relative strengths and weaknesses of human vs. AI on question answering, particularly with the new panoply of recently released LLMs, remains absent.

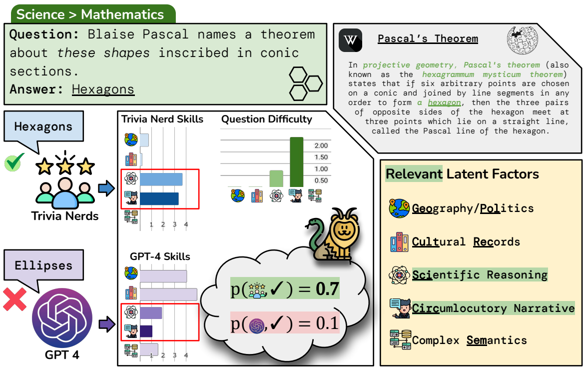

Figure 1: Response Correctness prediction using Agent skills and Question difficulty over relevant latent factors. We list the five latent factors that CAIMIRA discovers, and highlight the relevant ones (green), which contribute to estimating whether an agent will respond to the example question correctly. The agent skills over these relevant factors are highlighted in red boxes.

We seek to close that gap by contrasting problem-solving abilities of humans and AI for question answering (QA). We use a QA format (He et al., 2016; Rodriguez et al., 2019) specifically designed for effective comparison between QA agents (§ 2.1), that focus on rigorous trivia. The questions we choose are carefully crafted to probe the knowledge and reasoning the ability of human players and AI systems and expose the difference between them. Unlike Watson, rather than comparing one AI against two human teams on a couple of dozen questions, we compare ~ 30 AI systems against 155 humans on thousands of questions.

Our analysis of the QA agents is built upon improving item response theory (IRT, §2.2), a statistical framework that models the interaction between individuals and test items to assess their latent traits. First introduced in the field of Psychometrics (Santor and Ramsay, 1998), we use IRT to profile both the questions and agents. Classical IRT uses a one-

[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.

dimensional latent model that falls short of capturing the complexity inherent in response distributions that are best understood through a multidimensional lens. Additionally, its naïve multidimensional extension suffers from non-identifiability, where different combinations of difficulty and skills can yield identical responses. Furthermore, IRT identifies questions by unique indices, like `q35_2`, and not their textual content, and thus cannot extend to new questions with no agent response collected. To overcome these limitations, we propose a novel framework (§ 3): Content-aware, Identifiable, and Multidimensional Item Response Analysis (CAIMIRA, pronounced as ***Chimera***).

Applying CAIMIRA to responses collected from trivia players and a wide range of QA models over our questions (§ 4), we provide a thorough analysis of question and agent characteristics (§ 5). Our method uncovers five key latent factors (Figure 5), each encapsulating a distinct knowledge domain or reasoning skill, revealing specific facets of complexity in QA interactions.

Our findings show striking differences in humans and QA models' skills across these latent axes. Humans exhibit more consistent skills across all areas, outperforming AI in scientific reasoning and understanding indirect phrasing (circumlocution), reflecting their superior cognitive and interpretative abilities, Conversely, large-scale LLMs like GPT-4 and LLAMA-2-70B demonstrate superior ability in retrieving specific information about events and locations, often outdoing humans on questions loaded with entity-specific details—a feat we attribute to their extensive parametric memory. CAIMIRA also reveals questions that are easy for document recall but challenge most LLMs, and even humans to a certain degree, for answer recall. These adversarially crafted entity-rich questions utilize a lot of function words and complex semantics.

In conclusion, questions based on static knowledge pose less of an overall challenge than questions demanding deeper scientific understanding or nuanced language processing, suggesting that benchmarks focusing on scientific reasoning and linguistic intricacy are more discriminating in assessing QA agents' effectiveness.

2 Background and Preliminaries

This section describes the source of the human QA data (§ 2.1) and preliminaries of IRT and MIRT (§ 2.2), the foundation of CAIMIRA (§ 3).

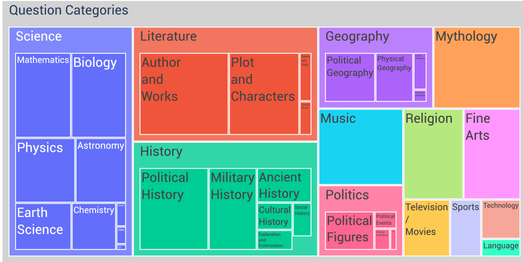

Figure 2: Distribution of question categories and subcategories over our dataset of 3042 questions.

2.1 QUIZBOWL: Where Trivia Nerds Practice

Our overarching goal is to identify similarities and differences between how systems and humans respond to questions. These questions must be *diverse*, less prone to ambiguity or false presuppositions, and designed to be challenging for humans so that we can draw conclusions about the strengths and weaknesses of agents without needing to "question the question" (Min et al., 2020; Yu et al., 2022). Following the categorization by Rogers et al. (2023), we prioritize "probing" questions that test depth over "information seeking" questions, focusing on trivia where responses from diverse competitive players are documented.

We use the "Protobowl" dataset (He et al., 2016), a dataset of trivia questions based on the Quizbowl (QB) QA setting (Boyd-Graber et al., 2012). Quizbowl, the source of questions for ProtoBowl, is a trivia game consisting of questions with sentence-clues decreasing in difficulty and culminating with a "giveaway" hint at the end of the question. 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 like history, science and literature.[2] (Figure 2)

We collect player 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 humans and AI systems by capturing their skills using Item Response Theory (IRT), a framework typically used to analyze human re-

[2] Appendix A provides further details into the QB dataset.

sponses (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 standardized tests for human subjects.

In the context of this work, IRT assumes (1) a set of question-answer pairs, (2) subjects spanning humans and QA systems, and (3) correctness rulings of their responses. The IRT objective is to predict the response correctness ($U_{i,j}$) based on the subject's skill s_i and the question's difficulty d_j, where i and j are the indices of the subject and question, respectively. The probability of response correctness, $p(U_{i,j} = 1)$, is modeled as $\sigma(s_i - d_j)$, where σ is the sigmoid function.

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

The learning objective here is to jointly model the skill and difficulty parameters that best estimate $p(U_{i,j})$ given the observed data. It is carried out using Bayesian inference assuming gaussian priors for the parameters.

Existing applications of IRT in NLP predominantly model item characteristics in one dimension. (Lalor et al., 2019). However, this approach assumes a linear hierarchy in difficulty and skill levels. For instance, if a history question q_h has higher difficulty than a science question q_s ($d_h > d_s$), the conventional IRT model assumes that agents who answer q_s correctly will also correctly answer q_h. The dimensional limitation of this model becomes particularly evident when considering the objective of distinguishing between human and computational agents in NLP tasks, necessitating a more nuanced and multi-dimensional approach.

Multidimensional Latent IRT (MIRT). To relax the monotonicity assumption, and model multi-factor characteristics, Chalmers (2012) proposes MIRT, which models two question characteristics, a scalar *difficulty* d_j, and an m-dimensional discriminability $\boldsymbol{\alpha_j}$ that interacts with the m-dimensional *skill* vector $\mathbf{s_i}$. The objective is then:

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

The discriminability $\boldsymbol{\alpha_j}$ captures how sensitively the correctness probability changes with each dimension of the agent skill $\mathbf{s_i}$. To mitigate overexpressibility, MIRT assumes $\boldsymbol{\alpha_j}$ to have a gamma prior, allowing only positive values. But, non-identifiability issues (Raue et al., 2009) persist.[3] A common practice of using hierarchical priors for resolving this makes optimization unstable in higher dimensions. Lastly, the model's exclusive dependence on question identifiers like `q31_2` over question *texts* hinders its ability to assess new questions without constant retraining, and treats questions as unrelated, risking noise interpretation as signal. The characteristics learnt this way do not identify the difference in the questions based on their content or source of the datasets (Rodriguez et al., 2022)

3 Bootstrapping IRT 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 introducing three key modifications: (i) a novel concept of relevance ($\mathbf{r_j}$) for each item j, (ii) zero-centered difficulty ($\mathbf{d_j}$), and (iii) learnable content-aware transformations ($\mathbf{W}_R$ and $\mathbf{W}_D$) from questions to their characteristics that can be applied to new questions. The CAIMIRA objective is:

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

where, $\mathbf{s_i} \in \mathbb{R}^m$ is agent skills,

and, $\mathbf{r_j}, \mathbf{d_j} \in \mathbb{R}^m$ are question relevance and difficulty resp.

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

Ideally, an *interpretable* item response analysis should include an item characteristic for each question that has the single responsibility of capturing how relevant each dimension is for estimating the likelihood of an agent correctly answering a particular question, $p(U_{i,j})$. We call this *relevance*.

To satisfy this, we decompose the combined information in MIRT's item characteristics, *discriminability* ($\boldsymbol{\alpha_j}$) and scalar difficulty (d_j) into more controlled m-dimensional characteristics, *relevance* ($\mathbf{r_j}$) and *difficulty* ($\mathbf{d_j}$), in CAIMIRA. Relevance $\mathbf{r_j}$ measures how differences between and agent skills and question difficulty ($\mathbf{s_i} - \mathbf{d_j}$), or *latent scores*, align across the dimensions (Eq 3), assigning each dimension (or, factor) a proportion ($\mathbf{r_{j,k}}$) to show its importance. To ensure clarity and prevent overlap with *difficulty*, $\mathbf{r_j}$ is defined

[3] Negative skill values ($\mathbf{s_i} < 0$) and their interaction with $\boldsymbol{\alpha_j} > 1$ could mimic similar likelihood estimates ($p(U_{i,j})$) as that of positive skills ($\mathbf{s_i} > 0$) with $\boldsymbol{\alpha_j} > 1$.

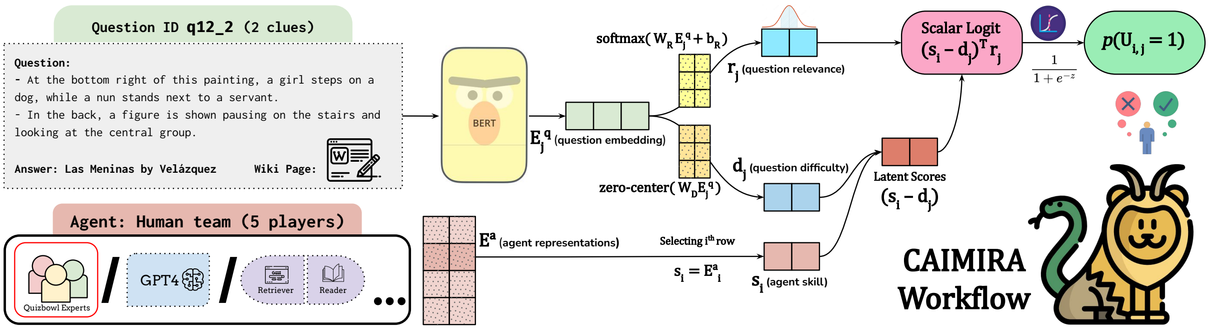

Figure 3: A pipeline of CAIMIRA. It predicts the probability of agent-i correctly answering question-j using a model in Eq. (3). Here, the question's raw relevance $\mathbf{r}'_\mathbf{j}$ and raw difficulty $\mathbf{d}_{:\mathbf{j}}$ are multidimensional and computed by learnt linear transformations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

as a probability distribution across the m dimensions. , guaranteeing that all values add up to one ($\sum_{k=1}^{m} \mathbf{r}_{\mathbf{j},\mathbf{k}} = 1$), and are non-negative.

For instance, in context of a quantum mechanics question, CAIMIRA assigns greater relevance to dimensions capturing physics knowledge and analytical reasoning, while downweighing unrelated dimensions like history or language. This targeted aggregation of differences across relevant dimensions ensures that the likelihood evaluation of an agent correctly answering the question, $p(U_{i,j} = 1 \mid \mathbf{s_i}, \mathbf{r_j}, \mathbf{d_j})$, is both precise and contextually appropriate.

Putting things together, $p(U_{i,j} = 1)$ is computed by aggregating the m-dimensional *latent scores* $(\mathbf{s_i} - \mathbf{d_j})$ to a scalar $(\mathbf{s_i} - \mathbf{d_j})^\intercal \mathbf{r_j}$ and applying the sigmoid function (σ) to it (Equation 3).

Connection to Topic Models This concept mirrors the mechanism in topic models, where documents are represented as mixtures of topics. Similarly, in CAIMIRA, questions are viewed as a mixtures of latent factors, or dimensions, with *relevance* $\mathbf{r_j}$ indicating the proportion of each dimension's contribution to the question. Just as topic models summarize a document's thematic structure by highlighting the most pertinent topics, CAIMIRA's relevance vector $\mathbf{r_j}$ distills the essential dimensions affecting question's difficulty and an agent's skill compatibility.

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

Aggregating *differences* between agent skills and question difficulty $(\mathbf{s_i} - \mathbf{d_j})$ across dimensions (Eq 3), leads to *non-unique* skill and difficulty values for same likelihood estimate $p(U_{i,j} = 1)$. We alleviate this non-identifiability issue by normalizing each question's **raw difficulty** $\mathbf{d}'_\mathbf{j}$ to have a zero mean for each dimension, maintaining the same correctness probability. This normalization constrains skill and difficulty ranges and enables comparisons across dimensions.

3.3 From MIRT to Content-Aware CAIMIRA

Unlike MIRT, CAIMIRA uses question text (content-aware) to compute characteristics and handle new questions at inference (cold-start friendly). Instead of learning the raw relevance ($\mathbf{r}'_\mathbf{j}$) and difficulty ($\mathbf{d}'_\mathbf{j}$) values for a question, it learns linear transforms ($\mathbf{W}_R$ and $\mathbf{W}_R$) from the question's embedding vector $\mathbf{E}^q_j$ to $\mathbf{r}'_\mathbf{j}$ and $\mathbf{d}'_\mathbf{j}$, which are then normalized to obtain $\mathbf{r_j}$ and $\mathbf{d_j}$. Mathematically,

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

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

where $\mathbf{W}_R, \mathbf{W}_D \in \mathbb{R}^{m \times n}$ and $\mathbf{b}_R \in \mathbb{R}^m$. These, along with the embedding matrix $\mathbf{E}^a$ of agent skills ($\mathbf{s_i} = \mathbf{E}^a_i$), are the parameters we train for CAIMIRA. The question embedding $\mathbf{E}^q_j$ is a high-dimensional representation of the question, which can be obtained using a pretrained transformer encoder like BERT, or a sparse BM25 representation.

Learning Objective. To regulate the question characteristics and agent skills learned by CAIMIRA, we adopt the Maximum A Posteriori (MAP) objective, combining the cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ (Equation 6) and regularization loss $\mathcal{L}_{\mathrm{reg}}$ (Equation 7). Specifically, the loss functions are defined as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i,j} \ell_{\text{CE}}(U_{i,j}, p(U_{i,j}=1)), \quad (6)$$

$$\mathcal{L}_{\text{reg}} = \lambda_d \sum_j \|\mathbf{d_j}\|_1 + \lambda_s \sum_i \|\mathbf{s_i}\|_1, \qquad (7)$$

$$\mathcal{L}_{\text{CAIMIRA}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{reg}}, \qquad (8)$$

where, $\ell_{CE}(x, y)$ represents the cross-entropy loss between the true label x and the predicted probability, y, $\|\cdot\|_1$ denotes the ℓ_1 norm, and λ_d and λ_s are the regularization hyperparameters.

4 Experimental Setup

This section describes how we collect responses from humans and QA systems, assess their answers, and analyze the latent traits learned by CAIMIRA from these responses.

Dataset Construction from Protobowl Logs. Protobowl questions are inherently multi-sentence constructs, with each sentence serving as a distinct clue about a specific entity or concept (the answer). Typically, a question has 4 clues on average. In our dataset, each item is formed by cumulatively adding clues from a Protobowl question, with the first item containing the initial clue and subsequent items incorporating an additional clue each.

Mapping Player Responses to Cumulative Clues. Player responses are mapped to these cumulative clue items to analyze the effectiveness of each clue set in eliciting correct answers. Responses to `q31` after only the first clue are recorded under `q31_1`, and responses after the second clue (which include the information from both clues) are recorded under `q31_2`, and so on. This mapping is further refined through a backfilling process. Because clues are meant to be progressively easier, we assume that a human who correctly answers a question at clue t, would also correctly answer the question at clue $t+1$. So, we mark those as correct as well. Similarly argument holds if humans answer incorrectly. With 3042 entries, our refined dataset and methodology provide a systematic analysis of how clue progression influences trivia response accuracy.

4.1 Human Agents

We aim to explore the complementarity between human and AI performance in answering questions. A key challenge in this investigation is the sparsity of comprehensive individual human data: most players only engage with a set of few dozen questions. To address this, we adopt a strategy of forming synthetic agents by grouping individual human players. This approach serves two primary purposes: it helps in accumulating a dataset where agents have attempted a substantial portion of the questions, and it mitigates the issue of non-representativeness of data from a few power users.

Group Formation and Decision Mechanism Our dataset comprises only five human players who have answered over 1500 questions each. While these "power users" are invaluable, relying solely on their data could skew the understanding of human-AI interaction, as they might not be representative of the broader player base. Therefore, we introduce the concept of "grouped human agents". Each grouped agent is a synthetic construct, representing an amalgamation of responses from multiple human players with similar skill levels. We group human players such that the overall coverage of questions attempted by the group is maximized. In cases where multiple players in a group answer the same question, we use a majority rule to determine the group's response. If no majority is reached, a response is sampled based on the votes.[4]

We consider group sizes of 1 (individual), 5, 10, and 15, creating five groups for each size, totaling 20 human agents spanning 155 distinct players.

4.2 AI Agents

To capture skill differentials across AI models and humans, and to learn about the advantages of various training and modeling techniques, we select a broad range of QA systems,[5] grouped as below:

Retrievers. These agents, indexing Wikipedia, use dense (e.g., CONTRIEVER (Izacard et al., 2021)) and sparse (e.g., BM25) methods to fetch the top k most relevant context documents to a query (where k = 1, 3, 5, 10). We call these context-retrievers. We also test a title-retriever, where only the document title(s) associated with the retrieved document(s) are considered as the answer predictions. Retrievers are evaluated on recall-based accuracy, with a point scored if the answer appears within retrieved documents for context-retrievers, or in the title for the title-retrievers.

Large Language Models (LLMs). We assess LLMs in a zero-shot setting, adhering to the stan-

[4]This method is a basic approach to represent group decision-making, acknowledging more complex dynamics for future research.

[5]Appendix B provides further details into model specs.

dard in-context learning practice (Brown et al., 2020), providing a task instruction followed by concatenated a single QA pair demonstration. These LLMs include base models (OPT (Zhang et al., 2022), GPT-Neo (Black et al., 2021) and Pythia (Biderman et al., 2023)), instruction-tuned models (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-2-70B (Touvron et al., 2023) and Falcon40B (Almazrouei et al., 2023)), and closed-sourced APIs (ChatGPT (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023)). In this work, we refer to the set of ChatGPT (or, GPT-3.5) and GPT-4 as GPT-3+. These models demonstrate a wide range of capabilities without being fine-tuned on our specific QA dataset.

Retriever-augmented Generative Models (RAG). Following the RAG paradigm (Lewis et al., 2020), we combine above defined retrievers with generative models for answer production, primarily using FlanT5-XL (Chung et al., 2022) with top 3 documents and exploring Flan-UL2 (Tay et al., 2022) for its larger receptive field to accommodate all ten.

Answer Match Equivalence. Traditional exact-match metric (Rajpurkar et al., 2016) often misses alternative answer that have different wordings or forms but the same semantic sense as the correct answer (Bulian et al., 2022). To better handle this, we adopt a fuzzy match evaluation using answer aliases (Si et al., 2021): if the character level matching rate between the predicted answer and the gold answer exceeds a certain threshold, the prediction is considered as correct. The threshold is tuned against human judgments on a small development set.

4.3 CAIMIRA Setup

We ablate to assess how number of latent dimensions, m, affect CAIMIRA's performance. Validation accuracy and loss plateaus beyond $m = 5$ (Figure 4), showing that it sufficiently captures question traits and agent skills. Thus, we train a 5-dimensional CAIMIRA model to learn the latent characteristics of questions and agents. SBERT (Reimers and Gurevych, 2019) provides with the question embeddings $\mathbf{E}_j^q$. We supplement SBERT's text input with both the answer and the first paragraph from its Wikipedia page, enhancing the contextual understanding of the question. The trainable parameters are fit using mini-

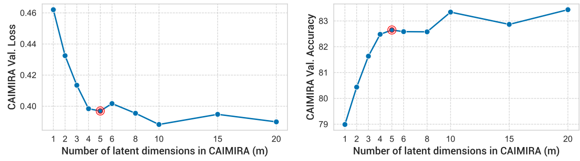

Figure 4: Ablation study showing CAIMIRA performance with varying latent dimensions m, indicating sufficiency at $m = 5$, beyond which gains are marginal.

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.

How do we interpret the latent factors? We want to study what nuances from question texts does CAIMIRA's 5-dimensional representations capture, and to what extent. For that, we use Logistic Regression as a supplemental interpretative tool to clarify the relationship between question texts and the characteristics identified by CAIMIRA.

We adopt the methodology from Gor et al. (2021), conducting a logistic regression analysis for each latent factor separately, using dimension-wise binary class labels assigned to every question according to its relevance value ($\mathbf{r_{jk}}$). For a dimension k, the class label is 1 if $\mathbf{r_{jk}} > 0.6$, and 0 otherwise. As input features, we use interpretable and hand-crafted features of the questions, e.g., topical question subcategories, clue counts, and a comprehensive set of linguistic features from Lee et al. (2021).[6] Thereby, we explain the latent factors in CAIMIRA by relating them to the logistic regression features with large (positive and negative) weights. Question categories are one-hot encoded; `c_plot_and_characters` is set to 1 for plot or character discussions, and 0 otherwise. The array of linguistic features span advanced semantic, discourse-based, and syntactic elements, providing a rich and multi-faceted representation of the questions. These are normalized to have zero mean and unit variance. Figure 5 lists the most contributing features for each dimension that are statistically significant (p-value < 0.05). To make the model fit (classification accuracy) comparable across dimensions, we incorporate class-balancing that maintains random guess accuracy for each dimension at 50%.

[6] Appendix C comprehensively lists all features we use.

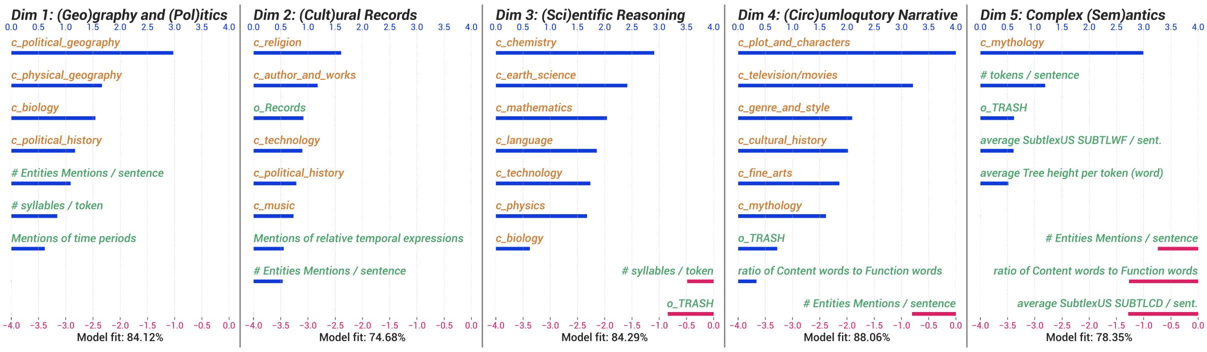

Figure 5: Interpretation of the five latent factors in CAIMIRA. We use Logistic Regression to predict the binary relevance label, $\mathbf{r_{jk}} > 0.6$, for each dimension k. We use question features that include topical categories (yellow) and linguistic properties (green). We report the classification accuracy and the statistically significant features. Coefficients are positive (blue bars) if the features positively affect classification, negative (red bars) otherwise. This demonstrates the efficacy of predicting the relevance from a question's SBERT embedding.

How do we interpret Question Difficulty? Our goal is to identify and categorize questions that are similar in terms of challenges they pose, to better understand their compositions and further create targeted benchmarks. For that, we inspect each question's *effective* difficulty. In the CAIMIRA objective (Eq 3), the effective contribution of the k-th dimension to the difficulty of question j is $\mathbf{r_{j,k}}\mathbf{d_{j,k}}$, we call this the *effective difficulty*, $\mathbf{d}^{(e)}_{\mathbf{j,k}}$. The aggregate of $\mathbf{d}^{(e)}_{\mathbf{j,k}}$ across all dimensions, $\mathbf{r_j}^\top\mathbf{d_j}$, quantifies a question's total difficulty, which also correlates with agents' average accuracy on question j. To achieve our goal, we use KMeans clustering to organize questions into twelve clusters based on their 5-dimensional effective difficulty $\mathbf{d}^{(e)}_{\mathbf{j}}$, and then examine the average *relevance* and *effective difficulty* within each cluster across dimensions (Figure 5).

5 Question and Agent Analysis

This section interprets CAIMIRA's latent factors using *relevance* (§ 5.1), and analyzes patterns in question difficulties and agent skills (§ 5.2).

5.1 Latent factors and Agent skills

The latent factors capture a variety of question styles and content, and the *relevance* of each factor is determined by the presence of specific linguistic and topical features in the questions (Figure 5). Human, context retrievers, and large scale LLMs exhibit stronger but complementary skills. While humans ace at science and questions with indirect phrasing with implicit context, GPT-4 excels at questions that have trigger phrases and are seeking time-specific information like geopolitical and record-setting events. Figure 6 compares the average skills of different agents by their categories across the five latent factors.

The first latent factor captures topics in *(Geo)graphy* and *(Pol)itics*. Questions associated have higher entity density, more polysyllabic words, and references to periods and locations. The second latent factor, *(Cult)ural Records*, reflects a question's focus on figures such as authors, composers, artists, and leaders. Questions often emphasize their record-setting achievements through terms like "most" and "first", and note a relative temporal context with words like "after", "before", and "recent". Large-scale LLMs show greater skills on these two dimensions.

The third latent factor, *(Sci)entific Reasoning*, highlights scientific phenomena and conceptual reasoning (e.g., "slope" in mathematics). These descriptive-styled questions, with an increased use of numbers, symbols, and multi-sense words and a deficit of entities pose a challenge to retrieval systems and smaller LLMs, while humans ace even the hardest of these questions. 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 next two latent factors focus on challenging and adversarially chosen question styles. The fourth one, though mostly related to literary works on surface, majorly captures *(circum)locution*, or indirect speech. Questions often narrate an event or describe characters typically from a fictional realm while deliberately avoiding direct references to named entities or key phrases (Fig 3). This style is a common source of difficulty in Quizbowl, especially for AI models. (Rodriguez et al., 2019). The

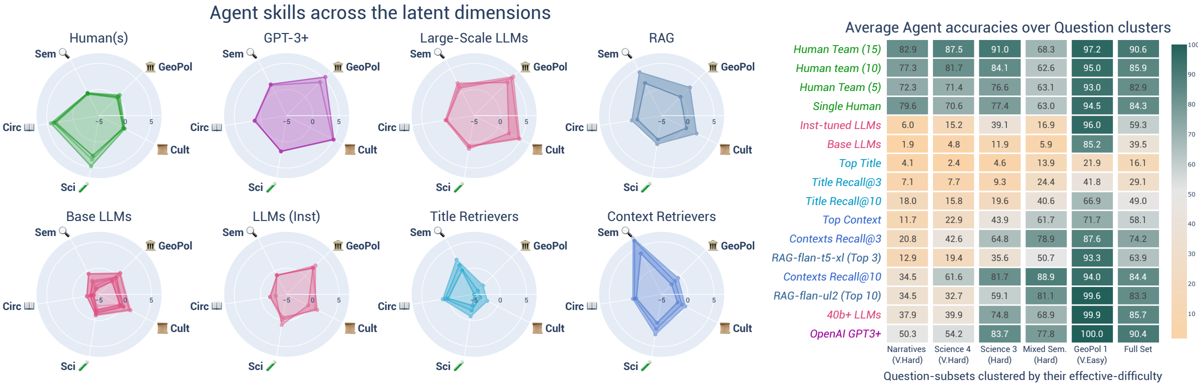

Figure 6: Left (radar plots) shows the average skills of our agents categories across our five latent factors (interpretations given in Figure 5). Right (heatmap) shows the accuracies of these agents types (rows) on questions clustered in their effective difficulty space (columns), first introduced in Figure 7.

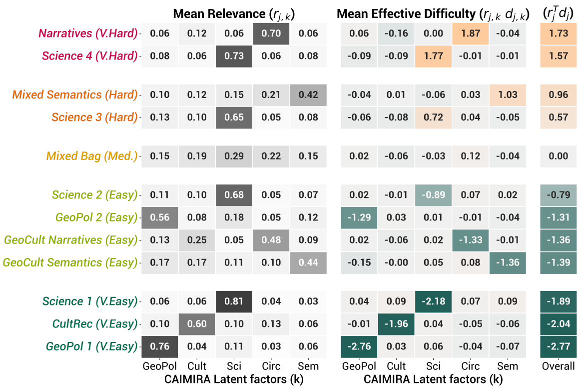

Figure 7: Heatmaps of relevance $\mathbf{r_{j,k}}$ and *effective difficulty* $\mathbf{r_{j,k}d_{j,k}}$ of question clusters (on effective difficulty) on the five latent factors (k) and the overall effective difficulty $\mathbf{r_j{}^\intercal d_j}$.

final latent factor, *Complex (Sem)antics*, pertains to questions on unusual events, termed TRASH (testing recall of strange happenings) in Quizbowl. These questions feature complex, detailed sentences with less common domain-specific words, which make them retriever-friendly (as shown in Figure 6) but hinder the extraction of answers by other agents due to intricate relationships among them. It appears that these questions were crafted based on how the Wikipedia articles about these events are written and the language used in them.

5.2 Which Questions are most difficult?

Figure 7 displays the *relevance* $\mathbf{r_{j,k}}$ and *effective difficulty* $\mathbf{d^{(e)}_{j,k}}$ of our twelve question clusters on the five latent dimensions, averaged within each cluster, and the heatmap in Figure 6 outlines the average accuracies of agents across these clusters, revealing notable distinctions: *Science 4* and and *Narratives* emerge as the most challenging categories, demonstrating high difficulty due to complex semantics, indirect phrasing and also mostly having a single clue. AI systems, including GPT-4, struggle with these, highlighting a marked disparity with human accuracy (Fig 6). Instruction-tuned LLMs outperform base ones in moderately difficult science questions (*Science 2*) with GPT-4 surpassing human teams of fewer than ten members. The distinction between easier and more difficult science questions, lies in their content: *Science 1* and *Science 2* have more clues, while *Science 3* and *Science 4* feature more numbers and symbols. *GeoPol 1* (Geography/Politics) and *Cultural Records* include the easiest questions; where base models lag slightly, whereas humans and GPT-4 nearly ace these factual queries with large number of clues, simple sentence structures and entity-rich content.

6 Related Work

Adoption of IRT in NLP. Current evaluation paradigms for machine and human QA inadequately segment datasets, treating questions as independent single transaction without assessing relative differences between the test set items. To remedy this, Lalor et al. (2019) propose adopting the IRT ranking method from educational testing as a novel evaluation framework for NLP. 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. Byrd and Srivastava (2022) further uses IRT to estimate question difficulty and model skills, and use question features to post-hoc predict question difficulty. Yet, existing studies are confined to a one-dimensional IRT models. Our research advances this domain by enhancing the learning method and capturing question traits that effectively differentiate human and AI QA abilities.

Ideal Point Models (IDP) IRT and IPM are two prominent statistical models used in different fields for distinct purposes. Both models deal with the analysis of preferences or abilities, but their applications and theoretical underpinnings show significant differences. IRT, used in educational assessments, gauges abilities from question responses, typically focusing on one-dimensional traits (De Ayala, 2013). Conversely, IPM, applied in political science, evaluates positions on spectra like political ideologies based on choices or votes (Clinton et al., 2004). Despite differences, both employ mathematically equivalent probabilistic methods to estimate the likelihood of a binary outcome—correctness in IRT, and votes in IDP, from a set of covariates, such as question difficulty or political ideology.

Human-AI Complementarity. Research in NLP has increasingly focused on augmenting human skills with language models, particularly in the areas like creative writing and question-answering. Studies have explored collaborative writing with LLMs, such as having human writers use GPT-3 for suggestions (Lee et al., 2022) or modifying user-selected text spans for enhanced descriptiveness (Padmakumar and He, 2021). For trivia, experts and novices have teamed up with AI (Feng and Boyd-Graber, 2018), and for information retrieval, humans used AI-generated queries to find answers (He et al., 2022) Our approach diverges by focusing modeling latent factors that best accentuate the distinct capabilities of trivia nerds and AI in QA. This strategy aims to identify the benchmarking methods for assessing and enhancing AI systems in subsequent work.

7 Conclusions

Our proposed CAIMIRA framework allows the discovery and interpretation of latent factors that best capture the nuances in question texts that are crucial in contrasting the strengths of human and AI for QA. We find a notable disparity in AI systems, like GPT-4, excelling at direct or context-rich queries and its struggles with subtle or indirect questions—domains where human acumen shines. This gap underscores the need for comprehensive datasets that more accurately assess a model's understanding of implicit contexts. Moreover, large language models (LLMs) resort to shortcuts when provided with adversarially crafted, semantically complex questions. These behaviors often lead to errors, despite apparent straightforward answers, emphasizing the need for future research to systematically categorize, identify and then mitigate shortcut-taking tendencies in these models. This becomes crucial as NLP evolves toward conversational agents and real-world problem-solving.

8 Limitations

Non-multilingual dataset Although there are QA datasets available spanning multiple languages, a majority of the AI systems that we use, with an exception of LLAMA-2-70B and GPT-4 fairly poorly on multilingual QA setting. Moreover, the there is no publicly available multilingual trivia with human responses and performance benchmarks.

Task-specific setup Although the QA task is a general task, and can encompass a wide variety of query based input/output tasks that can be assessed with binary correctness on an answer, there are no publicly available datasets that are not trivia based that have human responses in a competitive setting. Future work should focus on creating such datasets.

Lack of information on specific human players Because of the nature of the Protobowl platform that we used to collect the human response data, we do not have access to information about the specific human players to incorporate that into our analysis. Future work can focus on collecting such information whilst hiding the user identity.

Non-extensibliity of a trained CAIMIRA to a new agent. Unlike how CAIMIRA extended MIRT to model question characteristics as a function of question texts, and not just unique question identifiers, CAIMIRA is not extensible to a new agent without retraining the model. To make this possible for AI systems, future work can maintain a feature set that describes the specifications of an AI system that can include the model architecture, the training data, parameters, training strategies, etc, and have CAIMIRA learn a transformation from the feature set to agent skills. However, since this approach would require having a feature set for human players as well, which is not available, this approach is not feasible at the moment.

Static dense representation of from SBERT. In this work, we use a static dense representation of the question text from SBERT, instead of finetuning the model for adapting to CAIMIRA objective that learns representations from question text that best predicts the human response. This was out of the scope of this study. Future work can explore this direction using parameter efficient finetuning (PEFT) (Xu et al., 2023).

9 Ethical Considerations

In conducting this study, we adhered to strict ethical guidelines to ensure respect for privacy, obtaining informed consent from human participants and annonimization of their data. Our work complies with all relevant ethical standards, underscoring our commitment to ethical research practices in advancing NLP technologies. We utilized Copilot for coding and writing, and adhered to the highest standards of academic integrity and ethical conduct.

Regarding ethical considerations about running computationally expensive models, we acknowledge that the carbon footprint of training and running large-scale language models. In our study we only train a very small of order 25000 parameters, for 15 minutes of GPU time. We also use a pre-trained SBERT model for encoding the question text.

10 Acknowledgements

We thank UMD CLIP lab members: Neha Srikanth, Navita Goyal, Rupak Sarkar, along with the alumni: Pedro Rodriguez, Sweta Agrawal, and Chenglei Si for useful discussions and valuable feedback. We also thanks John Kirchenbauer for his suggestions on the toolings used for experimental evaluations. Finally, we express our gratitude to Flaticons for their extensive collection of icons which we utilize for making figures in this work.

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.

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*.

Matthew Byrd and Shashank Srivastava. 2022. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130, Dublin, Ireland. Association for Computational Linguistics.

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*.

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*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text.

A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929.

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*.

Pedro Rodriguez, Phu Mon Htut, John Lalor, and João Sedoc. 2022. Clustering examples in multi-dataset benchmarks with item response theory. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 100–112, Dublin, Ireland. Association for Computational Linguistics.

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*.

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 Schelten, 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*.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv: 2312.12148*.

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.

A Quizbowl Dataset

Quizbowl (Rodriguez et al., 2019), the source of questions for ProtoBowl, is a trivia game consisting of questions with clues decreasing in difficulty and culminating with a "giveaway" hint at the end of the question. The sequence of clues often reveals more information or helps disambiguate possible references and interpretations at each step. Figure 8 illustrates this structure with three example questions from different categories.

Question ID: q832_5 (Category: Religion)
This text was written down by Sahabas (sah-HAH-bahs) after the death of the leader that received it. The clarification of the meaning and significance of this document is the practice of tafsir (TAHFSEER). Its hundred and fourteen chapters are called suras (soor-AHS). It literally means "the recitation" and is said to have been revealed by Gabriel to Muhammad. For 10 points, what "divinely ordained" religious text is sacred to Muslims?
Answer: Piano / Pianoforte

Question ID: q622_3 (Category: Music)
Paul Wittgenstein ("VIT-gen-SHTINE") commissioned concertos for this instrument that used only the left hand. This instrument is said to have been invented by Bartolomeo Cristofori ("BAR-tow-lo- MAY-oh KRIS-tow-for-ee"). It was originally named for its ability to play both loud and soft sounds, which made it an improvement over the clavichord and harpsichord.
Answer: Piano / Pianoforte

Question ID: q2443_1 (Category: Science > Mathematics)
4 times the infinite sum one, minus one third, plus one fifth, minus one seventh, et cetera, equals this number.
Answer: pi / 3.14 / π

Figure 8: Example of QuizBowl questions for three different categories: Religion, Music and Mathematics, that illustrates the incremental nature of the questions.

Quizbowl naturally discriminates players' skills as players can **interrupt** questions to answer, and answering earlier is better.

In contrast to "all or nothing" QA, incremental QB questions help pinpoint the clues necessary for an agent a to answer question q by creating multiple opportunities for a to answer q. We achieve this by creating creating multiple entries for a single quizbowl question into our dataset. For instance, if a Quizbowl question **q622** has four clues in total, we create four entries, viz. **q622_1**, **q622_2**, **q622_3**, and **q622_4**, each corresponding to the question with first i clues, where $i \in \{1, 2, 3, 4\}$.

B QA Agents in our study

This section describes the QA agents used in our study, including the retrievers, LLMs, RAG models, and the prompts used to query them.

Retrievers as QA agents. Our retrievers, which index Wikipedia documents, respond with the top k documents (where k = 1, 3, 10) most relevant to the question. We employ two types of retrievers: dense and sparse. The dense retriever, CONTRIEVER (Izacard et al., 2021), is pretrained

Contexts Recall@10
- bm25_ctxs-recall@10
- contriever_ctxs-recall@10

Contexts Recall@3
- bm25_ctxs-recall@3
- contriever_ctxs-recall@3

Top Context
- bm25_ctxs-recall@1
- contriever_ctxs-recall@1

Figure 9: Agents we use in the Context Retrievers category.

Title Recall@10
- bm25_title-recall@10
- contriever_title-recall@10

Title Recall@3
- bm25_title-recall@3
- contriever_title-recall@3

Top Title
- bm25_title-recall@1
- contriever_title-recall@1

Figure 10: Agents we use in the Title Retrievers category.

via unsupervised contrastive learning on a mix of Wikipedia and CCNet data and 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 with index (Lin et al., 2021). We also test a title-retriever, assuming the document title is 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.,

40b+ LLMs

```
falcon-40b-instruct_1shot
falcon-40b_1shot
llama-2-70b_1shot
```

Base LLMs

```
gpt-neo-2.7B_1shot
opt-2.7b_1shot
opt-30b_1shot
pythia-12b-deduped_1shot
pythia-12b_1shot
pythia-2.8b-deduped_1shot
pythia-2.8b_1shot
pythia-6.9b-deduped_1shot
pythia-6.9b_1shot
```

Inst-tuned LLMs

```
flan-t5-xxl_1shot
flan-ul2_1shot
opt-iml-max-30b_1shot
```

Figure 11: Agents we use in the LLMs category.

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)

OpenAI GPT3+

```
openai-gpt-3.5-turbo_1shot
openai-gpt-4_1shot
```

Figure 12: Agents we use in the GPT-3+ category.

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. Figure 14 shows an example of the prompt used for these models.

RAG-flan-ul2 (Top 10)

```
rag-bm25_top10-flan-ul2
```

RAG-flan-t5-xl (Top 3)

```
rag-bm25_top3-T0pp-11b
rag-bm25_top3-flan-t5-xl
rag-contriever_top3-T0pp-11b
rag-contriever_top3-flan-t5-xl
```

Figure 13: Agents we use in the RAG category.

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. In our study, we mainly use FlanT5-XL (Chung et al., 2022) as the generator model, whose input context is limited to 512 tokens and composed of the top-3 documents by retriever. We also explore Flan-UL2 (Tay et al., 2022), an instruction-tuned UL2 with a 2048-token receptive field, to handle all the 10 documents. Figure 15 shows an example of the prompt used for RAG models.

```
You are a Quizbowl agent expert in Question Answering. Questions are in form of single or multiple clue(s) about a certain concept / entity. The following is a list of Quizbowl clues. Deduce the answer based on what the clues are describing, and answer the question in the form of a single word or a short phrase.

Question: { demonstration clues } What is being talked about here? Answer the question in a single word / short phrase.
Answer: { demonstration answer }

Question: { inference clues } What is being talked about here? Answer the question in a single word / short phrase.
Answer:
```

Figure 14: A condensed version of our prompt to Base models, Instruction-tuned models and Closed-source models (§ 4.2).

Answer Match Evaluation. Traditional exact-match metric often misses alternative answers that

```
You are a Quizbowl agent expert in Question Answering. Questions are
in form of single or multiple clue(s) about a certain concept / entity.
Answer the Quizbowl question by finding a short answer from the
reference documents listed below.

Documents:
{ Document 1 Title}: { Document 1 Content}
{ Document 2 Title}: { Document 2 Content}
. . .
{ Document k Title}: { Document k Content}

Question: { inference clues } What is being talked about here? Find the
answer from above documents and answer in a single word or a short
phrase.
Answer:
```

Figure 15: A condensed version of our prompt to our retriever-augmented generative (RAG) models (§ 4.2).

have different wordings or forms but the same semantic meaning as the correct answer (Bulian et al., 2022). To better handle this, we adopt a fuzzy match evaluation using multiple-answer aliases (Si et al., 2021): if the character level matching rate between the predicted answer and the gold answer exceeds a certain threshold, the prediction is considered as correct. The threshold is tuned against human judgments on a small development set.

C Question Features for Logistic Regression Study

This section describes the features used in the logistic regression study in § 4.3.

Question Category Features. These features are binary and indicate whether a question belongs to a specific category. These categories are the one highlighted in Figure 2. The categories are: `c_question_categories`, `c_fine_arts`, `c_cultural_geography`, `c_geography`, `c_physical_geography`, `c_political_geography`, `c_technical_geography`, `c_ancient_history`, `c_history`, `c_cultural_history`, `c_exploration_and_colonization`, `c_military_history`, `c_other`, `c_political_history`, `c_scientific_history`, `c_social_history`, `c_language`, `c_author_and_works`, `c_literature`, `c_genre_and_style`, `c_literary_terms`, `c_plot_and_characters`, `c_music`, `c_mythology`, `c_political_events`, `c_politics`, `c_political_figures`, `c_political_institutions`, `c_political_theory`, `c_religion`, `c_astronomy`, `c_science`, `c_biology`, `c_chemistry`, `c_earth_science`, `c_materials`, `c_mathematics`, `c_other`, `c_physics`, `c_scientific_history`, `c_sports`, `c_technology`, `c_television/movies`

Linguistic Features *LingFeat* is a Python research package designed for the extraction of various handcrafted linguistic features, positioning itself as a comprehensive NLP feature extraction tool. Currently, it is capable of extracting 255 linguistic features from English textual inputs. The features extracted by *LingFeat* span across five broad linguistic branches that Lee et al. (2021) details.

- **Advanced Semantic (AdSem):** Aims at measuring the complexity of meaning structures. Note: This feature is currently facing some operational issues, which are under investigation.
- **Semantic Richness, Noise, and Clarity:** Extracted from trained LDA models. The models are included and require no further training.
- **Discourse (Disco):** Focuses on measuring coherence and cohesion through entity counts, entity grid, and local coherence score.
- **Syntactic (Synta):** Evaluates the complexity of grammar and structure, including phrasal counts (e.g., Noun Phrase), part-of-speech counts, and tree structure.
- **Lexico Semantic (LxSem):** Measures word/phrasal-specific difficulty through metrics like type-token ratio, variation score (e.g., verb variation), age-of-acquisition, and SubtlexUS frequency.
- **Shallow Traditional (ShTra):** Encompasses traditional features/formulas for assessing text difficulty, such as basic average counts (words per sentence), Flesch-Kincaid Reading Ease, Smog, Gunning Fog, etc.

Time based features We create two time based feature, `t_range` and `t_range`. Both are binary features. `t_range` is 1 if the question was asked in the context of certain time period or a range, (e.g., *in the 20th century*, *in the 19th*), and 0 otherwise. `t_range` is 1 if the question refers to an event related to another event, (e.g., *after the fall of Rome*, *before the French Revolution*), and 0 otherwise.

Other features `o_TRASH` is 1 is the question enquires about specific events in pop culture category, and 0 otherwise. This feature reflects the TRASH category from Quizbowl. Similarly, `o_Records` is 1 if the question enquires about specific records through mention of superlative forms of words like "most recent", "best category", etc, and 0 otherwise. This feature reflects the Records category from Quizbowl.

D AI systems accuracies.

Narratives (V.Hard)

Answer: Nighthawks
Clues: This work was based on a real life location in Greenwich Village. It depicts a red-headed woman and two men in hats seated at a bar while being waited on by a man in a white hat.

Answer: matter
Clues: The density parameter for the non-relativistic form of this falls off with the cube of the scale factor.

Answer: Hermes
Clues: This deity led Perseus to the Gray Witches so he could kill Medusa.

Figure 16: Examples of questions from different clusters.

Science 4 (V.Hard)

Answer: (perfect) square numbers or perfect squares
Clues: The sum of the infinite sequence whose terms are the reciprocals of these numbers equals pi squared over 6.

Answer: Republic of Ireland
Clues: The head of the third largest bank in this country announced he had hidden 87 million Euros in loans from that bank. That announcement led to his arrest and the nationalization of that bank. In late November 2010, this country received an 85 billion Euro bailout from the EU.

Answer: WikiLeaks
Clues: A PowerPoint presentation released by this organization details how Bank of America plans to attack it.

Figure 17: Examples of questions from different clusters.

Mixed Semantics (Hard)

Answer: Saturn
Clues: Great White Spots are frequent storms on this planet.

Answer: Muammar al-Gaddafi
Clues: In 1969, this man seized power in a bloodless coup by overthrowing King Idris (EE-dreese). This author of The Green Book handed over the Lockerbie bombers after being visited by Nelson Mandela.

Answer: endoplasmic reticulum
Clues: One variant of this organelle ("OR-guh-NELL") is found in muscle cells and stores calcium. Like the Golgi body, it is composed of flattened sacks called cisternae ("SIS-ter-nay"). This set of tubes contains chaperone proteins, which help fold proteins.

Figure 18: Examples of questions from different clusters.

Science 3 (Hard)

Answer: Qur'an
Clues: Every chapter after the first chapter of this work is arranged from longest to shortest and all but one begins with the word "bismallah" (biss-MAH-lah).

Answer: 2
Clues: Euler characteristic of platonic solids have this value. This integer times pi gives the number of radians in the unit circle. Truth tables can evaluate to this many outputs. This value expressed in binary is 10 (ONE ZERO).

Answer: active transport
Clues: In nerve cells, this process is used to maintain the electrical membrane potential, and this process is also used to load sap into plant phloem. Most animal cells achieve this process with a sodium-potassium pump that is powered by ATP, while (*) phagocytosis of solid particles is another form of it. Used to move substances against the concentration gradient, for 10 points, name this transport process that requires energy.

Figure 19: Examples of questions from different clusters.

Mixed Bag (Med.)

Answer: Hermione Granger
Clues: This character was named after the wife of King Leontes in The Winter's Tale.

Answer: Theseus
Clues: This figure was nearly killed by his own father when Medea tricked the father into giving this figure a poisoned cup of wine. That cup was knocked away when this figure revealed a sword his father had hidden under a boulder with a pair of sandals.

Answer: Adam
Clues: According to the Koran, all angels, except Satan, prostrated themselves before this figure due to his knowledge. He was cursed to "eat bread until he returned to the ground."

Figure 20: Examples of questions from different clusters.

Science 2 (Easy)

Answer: friction
Clues: This force allows accelerated rolling motion down an incline by producing a net torque on the object. In general, this nonconservative force is equal to the normal force times mu, its namesake coefficient, and it converts kinetic energy into internal energy. For a given object, the kinetic variety is less than the static type. For 10 points, name this force between surfaces that opposes the motion of an object.

Answer: dark matter
Clues: It was the subject of a Scientific American special report dealing with Modified Newtonian Dynamics by Mordechai Milgrom ("MOR-de-kye MILL-grum"). This substance was first proposed in 1934 by Fritz Zwicky ("ZWICK-ee") to make up for "missing mass" in the universe. Its non-baryonic ("NON BARE- ee-on-ick") variety contains no mass.

Answer: tides
Clues: Arthur Doodson designed a machine for predicting the magnitude of these events. They occur in a cycle that includes "stand" periods followed by "slack water" periods. An unusually high concentration of dinoflagellates (DYE-no-FLADGE-ell-ates) can cause the "red" type. Weak versions of these events are known as "neap" ones and occur in the first and third quarters of the lunar cycle.

Figure 21: Examples of questions from different clusters.

GeoPol 2 (Easy)

Answer: State of the Vatican City
Clues: This country was officially recognized in the Lateran Treaties of 1929. It has extraterritorial authority over Castel Gandolfo.

Answer: Los Estados Unidos de México
Clues: A December 2012 agreement between this country's National Action, Democratic Revolution, and Institutional Revolutionary Parties led to constitutional amendments in 2013. The last of those parties is headed by (*) Enrique Peña Nieto [en-REE-kay PAY-nya nee-AY-toe], who replaced Felipe [fay-LEE-pay] Calderon as president. For 10 points, name this country that recently experienced an increase in drug-related violence and that shares a long border with the United States.

Answer: Hosni Mubarak
Clues: In 2003, this person warned that the Iraq War would create 100 bin Ladens. This person did not have a vice president until he appointed Omar Suleiman (OH-mar sue-LAY-mon) to that position. He originally declared he would not resign, which caused Tahrir Square to "[erupt] with anger," but reversed that decision the next day. For 10 points, name this former president and subject of mass uprisings in Egypt.

Figure 22: Examples of questions from different clusters.

GeoCult Narratives (Easy)

Answer: Nicolaus Copernicus
Clues: He published a then-controversial theory in "On the Revolutions of the Celestial Spheres," whose preface included a dedication to Pope Paul III so as to deflect controversy.

Answer: Osiris
Clues: This "Foremost of the Westerners" is linked with Serapis through the Apis bull. This son of Geb and Nut (NOOT) was cut into fourteen pieces that were scattered throughout the country by his brother.

Answer: The Hitchhiker's Guide to the Galaxy
Clues: In this novel, some mice fabricate a question that a super-computer was attempting to formulate, but it was destroyed minutes before the end of its 10 million year program.

Figure 23: Examples of questions from different clusters.

GeoCult Semantics (Easy)

Answer: King Arthur
Clues: A popular novel about this figure is T.H. White's The Once and Future King.

Answer: Antonio López de Santa Anna
Clues: This figure ordered the Goliad Massacre, and he was severely injured by French cannon fire at Veracruz during the Pastry War.

Answer: Aeneas
Clues: This man is told by the ghost of his wife Creusa to leave for Hesperia after carrying his father Anchises (ann-KYE-sees) and son Ascanius out of a besieged city. He visits the underworld with the help of a golden bough, on the advice of the Cumaean Sibyl. He duels Turnus for the hand of Lavinia. After this son of Venus leaves Carthage, Dido kills herself.

Figure 24: Examples of questions from different clusters.

Science 1 (V.Easy)

Answer: Moon
Clues: One theory of this entity's creation states that a Mars-sized body named Theia ("THEE-uh") collided with its parent planet. This object exhibits synchronous ("SIN-kro-nuss") rotation with its parent planet, and that rotation results in the namesake "dark side" of this object.

Answer: gravity
Clues: In standard units, this force's namesake constant equals 6.67 times ten to the negative eleventh power. This force's magnitude is inversely proportional to the square of the distance between two objects. On earth it causes objects to accelerate at 9.81 meters per second squared. It acts more strongly on objects of greater mass. For 10 points, name this fundamental force that causes objects to fall to the ground.

Answer: magnetism
Clues: Biot-Savart's Law gives the field of this type for a current carrying wire; the strength of that field is measured in Gausses and Teslas. There are para-, dia-, and ferro- forms of this phenomenon, the latter of which is expressed by metals such as nickel and iron. For 10 points, name this phenomenon whose field has both north and south poles, and which is often paired with electricity.

Figure 25: Examples of questions from different clusters.

CultRec (V.Easy)

Answer: The Crucible
Clues: Among those killed in this work is Giles Corey. Reverend Hale arrives to examine the unconscious Betty. This play sees Rebecca Nurse accused of killing seven of Goody Putnam's children, while Reverend Parris worries that his niece Abigail Williams will ruin his name. In the end, John Proctor refuses to make a false confession and is executed.

Answer: kinetic energy
Clues: A system's Lagrangian (lah-GRAN-jee-uhn) equals this quantity minus potential energy. This quantity can be found by dividing the square of an objects momentum by twice its mass. The change in this quantity for an object is equal to the net work done on the object. It equals one-half times mass times velocity squared. For 10 points, name this type of energy that objects possess because of motion.

Answer: M(aurits) C(ornelius) Escher
Clues: One of this man's works depicts his self-portrait in a glass ball situated on his hand. Another features two hands drawing each other into existence. This artist of the lithographs Hand with Reflecting Sphere and Drawing Hands created an ever-increasing stairway in Ascending and Descending, along with several tessellations. For 10 points, name this Dutch artist known for his fascination with optical illusions.

Figure 26: Examples of questions from different clusters.

GeoPol 1 (V.Easy)

Answer: The Canterbury Tales
Clues: One story in this work tells of the rooster Chauntecleer outsmarting a fox. Another story is about three rogues killing each other under an oak tree in a quest to find Death. In another story, a knight is forced to find out what women most desire; that story is told by the Wife of Bath. (*) Pilgrims on their way to visit an English cathedral city swap stories in, for 10 points, what collection by Geoffrey Chaucer?

Answer: France
Clues: One conflict in this country saw the Duke of Guise fight for the throne with two other men named Henry. This country signed the Evian Accords in 1962 with Algeria. In the 8th century, this was the site where Charles Martel was victorious at the Battle of Tours.

Answer: San Francisco, California
Clues: This city, home to the War Memorial Opera House, has such suburbs as Daly City.

Figure 27: Examples of questions from different clusters.

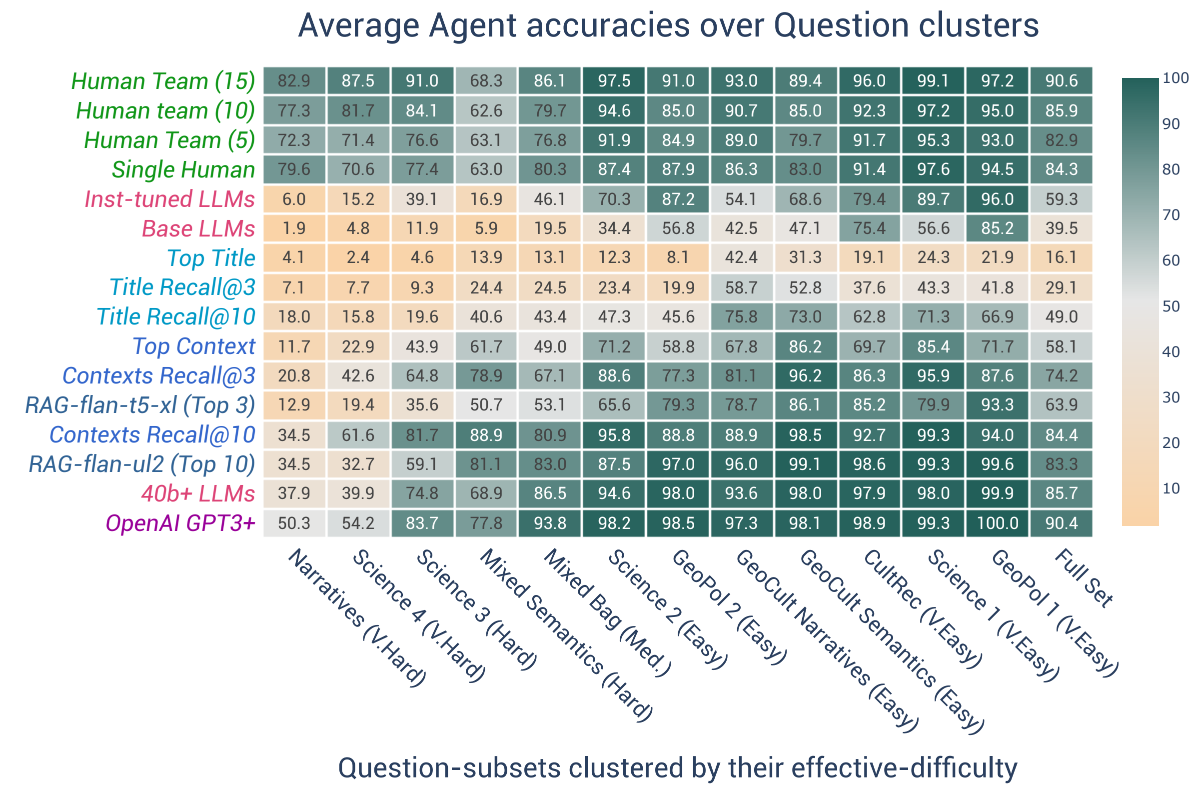

Figure 28: Full set of agent accuracies across all question clusters defined in section 5. We use the same color scheme as in Figure 6.