Do great minds think alike? Investigating Human-AI Complementarity in Question Answering with CAIMIRA

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

[1]University of Maryland [2]Microsoft Research
mgor@cs.umd.edu

Abstract

Recent advancements of large language models (LLMs) have led to claims of AI surpassing humans in natural language processing (NLP) tasks such as textual understanding and reasoning. This work investigates these assertions by introducing CAIMIRA, a novel framework rooted in item response theory (IRT) that enables quantitative assessment and comparison of problem-solving abilities of question-answering (QA) agents: humans and AI systems. Through analysis of over 300,000 responses from ~ 70 AI systems and 155 humans across thousands of quiz questions, CAIMIRA uncovers distinct proficiency patterns in knowledge domains and reasoning skills. Humans outperform AI systems in knowledge-grounded abductive and conceptual reasoning, while state-of-the-art LLMs like GPT-4-TURBO and LLAMA-3-70B show superior performance on targeted information retrieval and fact-based reasoning, particularly when information gaps are well-defined and addressable through pattern matching or data retrieval. These findings highlight the need for future QA tasks to focus on questions that challenge not only higher-order reasoning and scientific thinking, but also demand nuanced linguistic interpretation and cross-contextual knowledge application, helping advance AI developments that better emulate or complement human cognitive abilities in real-world problem-solving.

1 Introduction

The NLP community has focused on human behavior *emulation*, treating human performance as ceiling for models. However, the latest wave of LLMs has turned the discussion to supremacy: models are purportedly acing tests (Liu et al., 2023; Hendrycks et al., 2020) that many humans find challenging.[1]

[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.

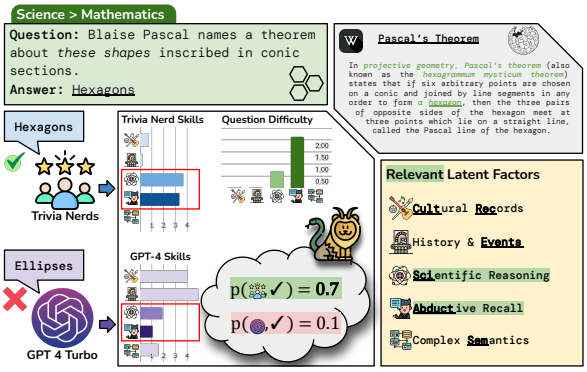

Figure 1: Response Correctness prediction using Agent skills and Question difficulty over relevant latent factors. We list the five latent factors that CAIMIRA discovers, and highlight the relevant ones (green), which contribute to estimating whether an agent will respond to the example question correctly.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

A notable 2010 example was IBM Watson's *tour de force* performance Ferrucci et al. (2010) on *Jeopardy!*. While Watson defeated the two humans on stage over a few dozen questions, a thorough, quantitative examination of the relative strengths and weaknesses of human vs. computer on question answering (QA), particularly with the new panoply of recent LLMs, remains absent.

To address this gap, we turn to Item Response Theory (IRT, §2.2), a statistical framework, originally developed in psychometrics (Santor and Ramsay, 1998), used for constructing effective standardized tests, by modeling the interaction between individuals and test items (questions). IRT is particularly suited for our analysis because it allows us to simultaneously assess the abilities of respondents (in our case, both humans and AI systems) and the characteristics of test items (our questions). This dual assessment is crucial for understanding the nuanced differences in performance between humans and AI systems across various types of questions.

Building upon IRT, we introduce CAIMIRA—

Content-aware, Identifiable, and Multidimensional Item Response Analysis (pronounced ***Chimera***)—a neural framework[2] that overcomes key challenges of applying IRT to QA. CAIMIRA uses question text to infer characteristics, enabling generalization to new questions without needing prior responses.

For our questions, we use a QA format (Boyd-Graber et al., 2012, QuizBowl) specifically designed for effective comparison between QA agents (§ 2.1). We then apply CAIMIRA (§ 5) to responses collected from 155 human trivia players, and a wide range (~ 70) of QA systems, over thousands of these carefully crafted questions that probe knowledge recall and reasoning capabilities. CAIMIRA uncovers latent aspects (Figure 5) that encapsulate different knowledge domains and reasoning skills, that best contrast agents' capabilities.

Humans and QA systems' skills are strikingly different across these latent axes (Figure 6). Human responses reflect their superior interpretative abilities, instinctive thinking, and cognitive flexibility. **This is particularly evident in questions demanding conceptual and knowledge-grounded abductive reasoning, characterized by indirect narrative references and ambiguous information gaps, where humans make intuitive leaps and draw connections that may not be immediately apparent.** Conversely, large-scale LLMs like GPT-4-TURBO and LLAMA-3-70B demonstrate superior ability in retrieving specific information about events and locations, outdoing humans on questions loaded with entity-specific details—a feat we attribute to their extensive parametric memory. CAIMIRA also reveals questions that, while easily matched to relevant documents by retrieval systems, challenge most LLMs in extracting the final answer. These questions feature complex sentence structures and semantic relationships, that turn simple information retrieval into demanding reading comprehension.

In conclusion, this study provides insights into the strengths and weaknesses of human and AI question answering, laying the groundwork for future AI developments that better emulate or complement human cognitive abilities. In doing so, it underscores the need for sophisticated benchmarks to controllably distinguish between proficient and less capable QA systems, especially in areas demanding deeper, conceptual, and linguistic understanding.

[2]The implementation can be found at https://github.com/maharshi95/neural-irt

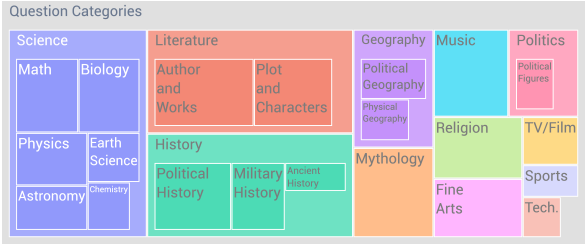

Figure 2: Distribution of question categories and subcategories over our dataset of 3042 questions.

2 Background and Preliminaries

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

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 false presuppositions, and designed to be challenging for humans, enabling us to 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 focus on depth-testing "probing" questions over "information seeking" ones. This approach aligns with the Manchester paradigm outlined by Rodriguez and Boyd-Graber (2021), which highlights the significance of research agendas in the development of human-like, intelligent QA systems. More importantly, we need questions with many examples of diverse human answers. While humans may not answer Google queries (Kwiatkowski et al., 2019) for fun, they do answer trivia questions as a hobby or to prepare for trivia competitions. Hence, we use the "Protobowl" (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. 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[3] (Figure 2).

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

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 used to understand question quality and participant strengths, by analyzing 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 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) binary 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.

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

The learning objective is to model skill and difficulty parameters that best fit assumed priors, given observed response data, typically using Bayesian inference. Existing IRT applications in NLP often model item characteristics in one dimension (Lalor et al., 2019), assuming a linear hierarchy in difficulty and skill levels. This approach is limiting when distinguishing between agents in NLP tasks. For example, if a history question q_h is found to be more difficult than a science question q_s ($d_h > d_s$), the model asserts that agents correctly answering q_h also correctly answer q_s, and vice versa.

Multidimensional Latent IRT (MIRT). To relax the monotonicity assumption and model multifactor characteristics, MIRT was developed (Reckase, 2006; Chalmers, 2012). It 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 skill value $\mathbf{s_{i,k}}$ corresponds to the agent's expertise on the k^{th} *latent aspect*. The objective then becomes:

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

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.[4] A common practice of using hierarchical priors for resolving this makes optimization unstable for higher dimensions. Lastly, the model's exclusive dependence on question identifiers (`q31_2`) treats questions as unrelated and hinders generalization. The characteristics learned this way do not identify the difference in the questions based on their content (Rodriguez et al., 2022)

3 Bootstrapping IRT with CAIMIRA

We propose CAIMIRA—Content-aware, Identifiable, and Multidimensional Item Response Analysis, an IRT framework 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 (f_R and f_D) that produce $\mathbf{r_j}$ and $\mathbf{d_j}$ from the raw questions. These enable CAIMIRA to provide interpretable and identifiable results, and handle new questions without prior response data. The response prediction model, the probability of agent i correctly answering question j, for an m-dimensional CAIMIRA, is given by Equation 3.

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

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* $\mathbf{r_j}$

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

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 m-dimensions (Eq 3), assigning each dimension (or, latent aspect) a proportion ($\mathbf{r_{j,k}}$) to show its importance. To ensure clarity and prevent overlap with *difficulty*, $\mathbf{r_j}$ is defined as a probability distribution across the m dimensions. For instance, for a Thermodynamics question, CAIMIRA assigns greater

[4] 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 $\boldsymbol{\alpha_j} > 1$.

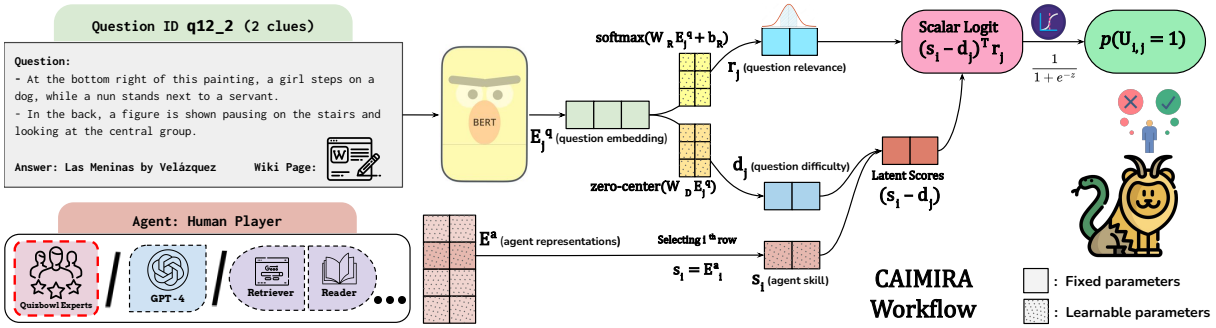

Figure 3: The CAIMIRA workflow. It predicts the probability of agent-i correctly answering question-j using a model in Eq. (3). Here, the question's raw relevance $\mathbf{r}'_\mathbf{j}$ and raw difficulty $\mathbf{d}_{;\mathbf{j}}$ are multidimensional and computed by learnt linear transformations 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).

relevance to dimensions capturing physics knowledge and analytical reasoning, down weighing unrelated dimensions like history or language. This targeted aggregation of differences across relevant dimensions ensures that the likelihood estimate $p(U_{i,j} = 1 \,|\, \mathbf{s_i}, \mathbf{r_j}, \mathbf{d_j})$, is both precise and contextually appropriate.

Connection to Topic Models This admixture mirrors the per-document allocation in topic models; in CAIMIRA, questions are admixtures of latent aspects, or dimensions, with *relevance* $\mathbf{r_j}$ indicating each dimension's contribution to the question.

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 (Equation 7). This normalization constrains skill and difficulty ranges and enables comparisons across dimensions.

3.3 Content-Aware Transformations

CAIMIRA improves upon MIRT by incorporating question content, enabling CAIMIRA to compute characteristics for new questions without requiring prior response data, making it "cold-start friendly". At its core, CAIMIRA maps question text into *relevance* and *difficulty* values using learnable functions, $f_R, f_D : Q \to \mathbb{R}^m$, transforming a question q_j from the space of question texts Q into raw relevance $(\mathbf{r}'_\mathbf{j})$ and raw difficulty $(\mathbf{d}'_\mathbf{j})$ vectors (Figure 3). These are modeled as linear transformations over a pre-trained embedder $f_E : Q \to \mathbb{R}^n$ (e.g., BERT), which represents $q_j \in Q$ in an n-dimensional space as an embedding $\mathbf{e_j}$:

$$\mathbf{e_j} := f_E(q_j) = \texttt{BERT}(q_j), \tag{4}$$

$$\mathbf{r}'_\mathbf{j} := f_R(q_j) = \mathbf{W}_R\, \mathbf{e_j} + \mathbf{b}_R, \tag{5}$$

$$\mathbf{d}'_\mathbf{j} := f_D(q_j) = \mathbf{W}_D\, \mathbf{e_j} \tag{6}$$

where $\mathbf{W}_R, \mathbf{W}_D \in \mathbb{R}^{m \times n}$ and $\mathbf{b}_R \in \mathbb{R}^m$ are the parameters of the linear transformations.[5] The raw values are then normalized to obtain final relevance $(\mathbf{r_j})$ and difficulty $(\mathbf{d_j})$ values:

$$\mathbf{r_j} := \text{softmax}(\mathbf{r}'_\mathbf{j}), \quad \mathbf{d_j} := \mathbf{d}'_\mathbf{j} - \frac{1}{n_q} \sum_{j=1}^{n_q} \mathbf{d}'_\mathbf{j}, \tag{7}$$

where n_q is the number of questions in the dataset. softmax normalization for relevance ensures that the values sum to 1 across m-dimensions, reflecting the relative importance of each latent aspect.

Agent Skills. CAIMIRA learns an agent skill embedding matrix $\mathbf{E}^a \in \mathbb{R}^{n_a \times m}$, where n_a is the number of agents, and the skill vector for agent i is the i^{th} row of this matrix:

$$\mathbf{s_i} = \mathbf{E}^a_i \tag{8}$$

This approach allows CAIMIRA to learn a compact representation of each agent's skills and question characteristics (difficulty and relevance), across m dimensions, which can be directly used in the response prediction model (Equation 3).

Learning Objective. To optimize CAIMIRA's parameters (Θ), which include the agent skill embedding matrix $\mathbf{E}^a$ and the linear transformation

[5] We skip the bias term for $\mathbf{d}'_\mathbf{j}$ since it is mean-centered.

parameters $\mathbf{b}_R$, $\mathbf{W}_R$ and $\mathbf{W}_D$, we use *maximum a posteriori* estimate (MAP) based loss, which imposes implicit priors on the question characteristics and agent skills. This combines a cross-entropy loss $\mathcal{L}_{\text{CE}}$ (Eq 9) with regularization terms (Eq 10):

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i,j} \ell_{\text{CE}}(U_{i,j}, p(U_{i,j}=1)), \quad (9)$$

$$\mathcal{L}_{\text{reg}} = \lambda_d \sum_j \|\mathbf{d_j}\|_1 + \lambda_s \sum_i \|\mathbf{s_i}\|_1, \quad (10)$$

where $\ell_{CE}(x, y)$ is the cross-entropy loss between the true label x and the predicted probability in Eq. (3), y. $\|\cdot\|_1$ denotes the ℓ_1 norm, and λ_d and λ_s are the regularization hyperparameters. Finally,

$$\mathcal{L}_{\text{CAIMIRA}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{reg}}, \quad (11)$$

$$\Theta_{\text{CAIMIRA}} \triangleq \arg\min_{\Theta} \mathcal{L}_{\text{CAIMIRA}} \quad (12)$$

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.

Protobowl Logs. We collect player logs from the "Protobowl" platform over QB questions spanning various categories. (Figure 2) Player logs record question metadata, including category (e.g. History), time taken to answer the question, answer string, and the correctness ruling by the platform. The best players have deep knowledge and excellent lateral thinking skills (Jennings, 2006).

Constructing QA Dataset. QB questions are inherently multi-sentence (typically five) with each sentence serving as a distinct clue for the answer. In our dataset, each item is formed by cumulatively adding clues from a QB question, with the first item containing the initial clue and subsequent items incorporating an additional clue each; i.e., the first item consists of only the first clue, the second item comprises the first two clues together, and so on. This cumulative clue addition provides insight into how progressively revealing information affects agents' response accuracy.

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 player who correctly answers a question at clue t, would also correctly answer the question at clue $t' > t$. So, we mark those as correct as well. An analogous argument holds for $t' < t$ when humans answer incorrectly. Consequently, we collect a total of 3042 entries in our refined dataset.[6]

4.1 Human Agents

In exploring the complementary QA abilities of human and AI, a key challenge is the sparsity of individual human data: most players only engage with a set of few dozen questions. To address this, we form synthetic human 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 "grouped human agents". Each grouped agent is a synthetic construct, amalgamating 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.[7]

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. Our human participants, all fluent in US English, are experienced Quiz Bowl players. While this sample may not encompass the full diversity of the broader population, their expertise in trivia games, particularly in Quiz Bowl, allows us to contrast the

[6] The dataset is available on the HuggingFace platform as `mgor/protobowl-11-13`.

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

nuanced skill sets of seasoned Quiz Bowl enthusiasts with the capabilities of our AI systems.

4.2 AI Agents

To capture skill differentials across AI models and humans and to learn the effects of various training and modeling techniques, we select a broad range of QA systems,[8] grouped as below:

Retrievers. These agents, indexing Wikipedia, use sparse (e.g., BM25), and dense—GRIT-LM (Muennighoff et al., 2024) and CONTRIEVER (Izacard et al., 2021)—methods to fetch the 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 title(s) associated with the retrieved document(s) are answer predictions. Retrievers are evaluated on recall, 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 zero-shot in-context learning (Brown et al., 2020), providing a task instruction followed by 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 like LLAMA-3-70B (Touvron et al., 2023), Falcon40B (Almazrouei et al., 2023), Cohere's CMD-R+[9] and Mixtral 8x7b (Jiang et al., 2024), and closed-sourced APIs such as GPT-4O, GPT-4-TURBO (OpenAI, 2023) and Gemini-family (Team et al., 2024).

Retriever-augmented Generative Models (RAG). 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), and CMD-R+ to accommodate all ten.

Answer Match Equivalence. Traditional exact-match (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. We tuned the threshold against human judgments on a small dev set.

[8] Appendix C provides further details into model specs.

[9] https://docs.cohere.com/docs/command-r-plus

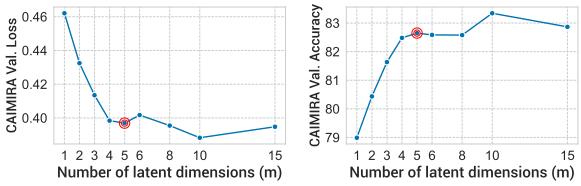

Figure 4: Ablation study showing CAIMIRA performance with varying latent dimensions m, indicating sufficiency at $m = 5$, beyond which gains are marginal.

4.3 CAIMIRA Setup

We ablate the number of latent dimensions, m. Validation loss plateaus beyond $m = 5$ (Fig 4). We thus train a 5-dimensional CAIMIRA model using `all-mpnet-base-v2`, an SBERT variant (Reimers and Gurevych, 2019) as the question embedder f_E. To capture information gaps between questions and answers, we supplement SBERT's text input with both the answer and it's Wikipedia page summary. We minimize $\mathcal{L}_{\text{CAIMIRA}}$ (Equation 11) using Adam optimizer (Kingma and Ba, 2014), with learning rate 0.005, batch size 512, and $\lambda_d = \lambda_s = 1e-5$.

Interpreting Latent Aspects. To study the latent dimensions of CAIMIRA, we use Logistic Regression as a supplemental interpretative tool. We build upon Benedetto et al. (2020), which uses Linear Regression to post-hoc explain the latent item difficulty parameters, and follow Gor et al. (2021) to interpret the latent relevance dimensions using logistic regression. For each latent dimension (k), Logistic Regression predicts if the relevance $\mathbf{r_{jk}}$ is greater than 0.6 as a function of interpretable features extracted from the questions. These features span topical question subcategories, clue counts, temporal expression mentions, question similarity with corresponding Wikipedia pages (WikiMatchScore), and linguistic features from Lee et al. (2021).[10] Thereby, we explain CAIMIRA's latent dimensions by relating them to the logistic regression features with large (positive and negative) coefficients. Topical features are one-hot encoded; `c_music` is set to 1 for music related question, and 0 otherwise. The linguistics features span advanced

[10] Appendix D lists all features we use.

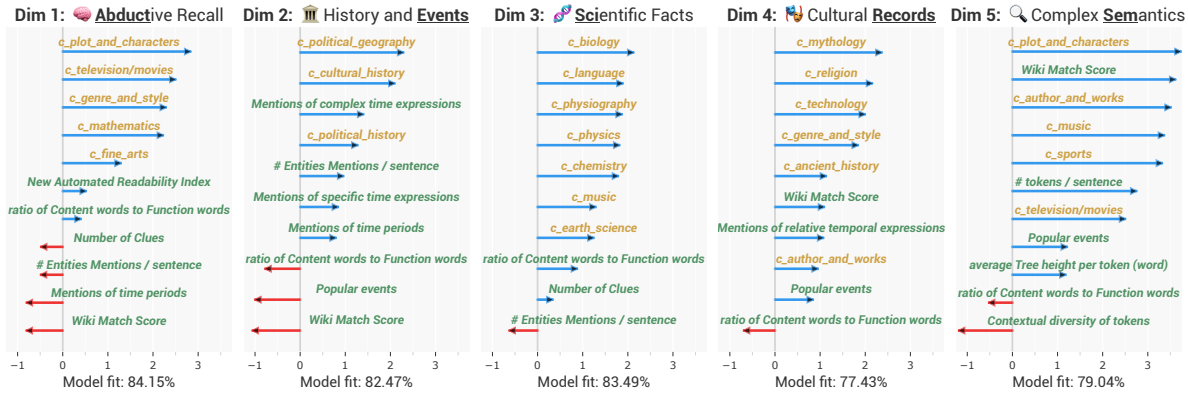

Figure 5: **Interpretation of the five latent dimensions in CAIMIRA.** We use Logistic Regression to predict the binary relevance label, $\mathbf{r_{jk}} > 0.6$, for each dimension k. For question features, we use **topical categories** and **linguistic properties**. We report the classification accuracy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

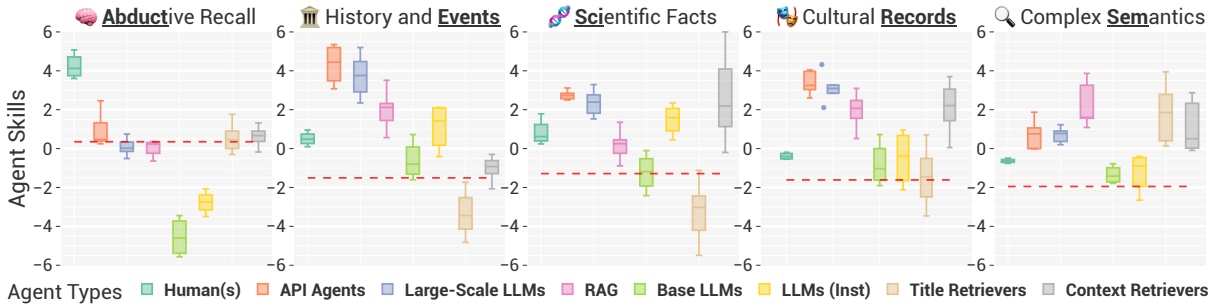

Figure 6: Distribution of skills grouped by agent type across the five latent dimensions of CAIMIRA. Interpretations given in Figure 5. The red dashed line indicates the mean *effective difficulty* of each dimension (Equation 13).

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, statistically significant features for each dimension (p-value < 0.05). To make the learned coefficients comparable across dimensions, we incorporate class-balancing maintaining the random guess accuracy for each dimension at 50%.

5 Question and Agent Analysis

This section interprets the latent aspects of CAIMIRA, emphasizing their role in differentiating agent skills. It also examines the patterns of question difficulty and agent performance.

5.1 Latent aspects and Agent skills

CAIMIRA uncovers five latent aspects, each capturing distinct question styles and content, determined by specific linguistic and topical features (Figure 5). These aspects highlight varying agent skills across the latent dimensions (Figure 6). In naming and interpreting these aspects, we draw on educational assessment frameworks, particularly Bloom's Taxonomy (Anderson and Krathwohl, 2001), which emphasizes the stages of knowledge recall, comprehension, and application—skills central to the Quizbowl dataset.

Abductive Recall. The first aspect captures a cognitive process that combines elements of inferential reasoning with targeted knowledge retrieval. It requires bridging indirect clues and vague references to formulate the information gap, and recalling specific entities to fill the gap. This distinguishes it from purely creative and commonsense-based abductive reasoning tasks in linguistics literature (Bhagavatula et al., 2019; Shi et al., 2024). We term this aspect "abductive recall" to highlight the interplay between hypothesis generation and gap resolution through targeted fact retrieval. Questions often narrate events and describe characters from a fictional realm while deliberately avoiding direct references to named entities or key phrases (Example in Fig 3). A low WikiMatchScore—semantic

overlap between questions and their associated Wikipedia pages—combined with the absence of entities and key phrases, indicate a significant information gap that necessitates not just multi-hop reasoning skills to bridge the contextual divide, but also deducing relevant involved entities from the narrative. Humans excel at these questions, surpassing GPT-4-TURBO by leveraging intuition to connect abstract clues to specific entities, while most AI models struggle.

History and Events. In contrast, the second dimension involves historically grounded questions, where the information gap is clearer, though the queries are complex. These questions challenge participants to synthesize multiple pieces of information and infer connections between events. For e.g, `"This man was killed by a crossbow bolt while besieging the castle Charlus-Chabrol"`, requires identifying both the event and the historical figure. While these questions still feature lower WikiMatchScores, the gap is more structured, centering around entity relations like events, people, and places. Bigger LLMs excel in this category, often outperforming humans and retrievers, suggesting effective recall and application of historical information through their parametric memory.

Scientific Facts. This aspect focuses on domain-specific conceptual knowledge, often featuring questions from scientific domains. Retrieval-based systems fare well when allowed to retrieve sufficient documents (Figure 7). Notably, these questions, along with history-related ones, best differentiate instruction-tuned LLMs from base models, with the former outperforming the latter. Humans and large-scale LLMs excel in this category, as do closed-source systems like GPT-4-TURBO.

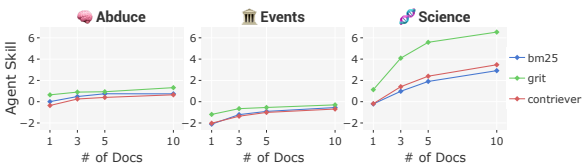

Figure 7: Variation in Context Retriever skills across latent dimensions as the number of retrieved documents (top-k) increases, showing that a system which retrieves more documents can achieve higher skills in *Science*, but not on *Abduction* and *Events*.

Cultural Records. This aspect represents questions focusing on prominent figures such as authors, composers, artists, and leaders, asked in the style of "who did what", testing direct knowledge recall

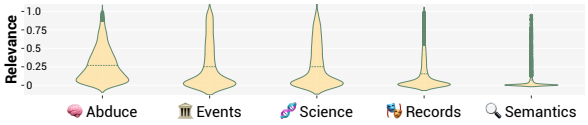

Figure 8: Distribution of *relevance* ($r_{j,k}$) scores across CAIMIRA's five latent dimensions. Cultural Records and Complex Semantics are not as representative of the dataset, as the first three.

of well-known facts and making them relatively easy and accessible (high WikiMatchScore).

Complex Semantics. The final aspect pertains to questions about popular events, featuring complex semantic relationships and detailed sentences with less common, domain-specific keywords. Despite their intricacy, they are particularly retriever-friendly due to high WikiMatchScores, indicating a significant overlap with relevant source documents. The most prominent fact about the answer is directly mentioned in both the question and the document, enabling retrievers to locate correct documents. However, agents without retrieval abilities, or large parametric memories, struggle.

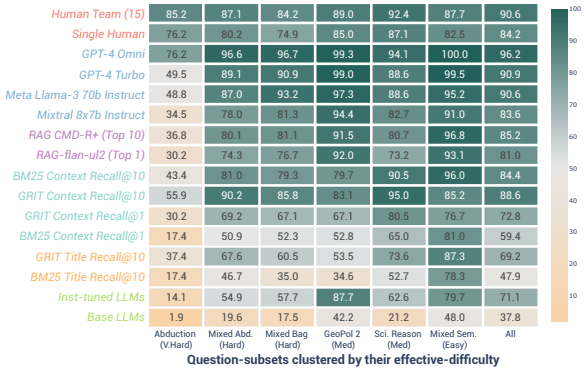

Figure 9: Agent accuracies on various dataset slices.

5.2 Which Questions are most difficult?

To identify groups of questions that present different challenges, we analyze each question's *effective difficulty*, denoted as $\mathbf{d}^{(\mathbf{e})}_{\mathbf{j,k}}$. This metric represents the contribution of the k-th latent aspect to the difficulty of question j, calculated as $\mathbf{r_{j,k}}\mathbf{d_{j,k}}$ according to Equation 3. We cluster questions into twelve groups using KMeans on their 5-dimensional effective difficulty $\mathbf{d}^{(\mathbf{e})}_{\mathbf{j}}$, then analyze *mean relevance* and *mean effective difficulty* per cluster across dimensions (Fig 10, full set in Appendix E). The mean effective difficulty $\mathbf{d}^{(\mathbf{e})}_{\mathbf{D},\mu_\mathbf{k}}$ on the dimension k for a question set D is calculated as a weighted mean of the effective difficulty scores over the ques-

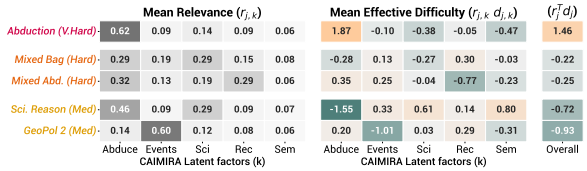

Figure 10: Heatmaps of mean relevance $\mathbf{r_{j,k}}$ and *mean effective difficulty* $\mathbf{d}^{(\mathbf{e})}_{\mathbf{D},\mu_\mathbf{k}}$ of selected question clusters (on effective difficulty) across the five latent factors (k).

tions in D, normalized by the total relevance.

$$\mathbf{d}^{(\mathbf{e})}_{\mathbf{D},\mu_\mathbf{k}} = \frac{\sum_{j \in D} \mathbf{r_{j,k}} \mathbf{d_{j,k}}}{\sum_{j \in D} \mathbf{r_{j,k}}} \tag{13}$$

Abduction (V.Hard) and *Mixed Bag* 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-TURBO, struggle with these, highlighting a marked disparity with human accuracy (Fig 9). Instruction-tuned LLMs outperform base ones in moderately difficult science questions, with GPT-4O surpassing single human players. A common trend we observe is that for each latent factor, questions tend to have higher difficulty when they have fewer clues, and lower WikiMatchScore.

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

CAIMIRA enables discovery and interpretation of latent aspects in QA datasets that highlight the skills of various QA agents. On contrasting AI systems with humans, we find notable disparities: systems like GPT-4-TURBO and Gemini Pro excel at direct, context-rich queries that require connecting events and figures, but struggle with indirectly phrased questions lacking explicit entity references—domains where human acumen shines. Although GPT-4-TURBO matches individual human performance on complex knowledge-intensive abductive reasoning tasks, we caution against interpreting this as indicative of superhuman abilities. Given that the quiz questions that Protobowl is based off have been publicly available since 2011, and that these models' training data is not fully known, accurately assessing the reason for their

near-perfect performance is challenging. Future research should aim to develop stronger and innovative evaluations that better gauge AI systems' ability to understand implicit contexts, and systematically contrast their skills with those of humans. Lastly, this work opens up new avenues for research on estimating agent skills that can be combined to assess multi-agent systems and collaborations, which becomes crucial as NLP evolves toward conversational agents and real-world problem-solving.

8 Limitations

Dataset and Task Limitations Our study faces constraints related to dataset and task setup: (1) Limited language diversity: Our English-only dataset restricts generalizability to other languages. (2) Lack of diverse task types: We rely solely on trivia-based questions, lacking non-trivia datasets with human responses in competitive settings. (3) Absence of multilingual trivia benchmarks: We lack multilingual trivia datasets with human responses and performance benchmarks. Future work should address these by creating datasets that include non-trivia tasks, multiple languages, and human responses, offering a more comprehensive understanding of human and AI performance across diverse linguistic and task environments.

Challenges in interpreting near-perfect scores While models like GPT-4-TURBO match or exceed individual humans on complex tasks, caution is needed when interpreting these results as superhuman. Quiz questions in our Protobowl-based dataset have been public since 2011, and the models' full training data is unknown. This makes it difficult to determine if their near-perfect performance stems from genuine reasoning or exposure to specific questions during pre-training. genuine reasoning or exposure to specific questions during pre-training. This limitation highlights the need for more robust evaluation methods to accurately assess AI systems' understanding and reasoning abilities compared to humans.

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 AI systems. 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 representation 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 GitHub Copilot for low level coding and writing assistance—reimplementing plotting codes, as well as editing the prose in this document to improve readability and conciseness.

Regarding ethical considerations about running computationally expensive models, we acknowledge 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 20 minutes of single A4000 GPU time. We also use a pre-trained SBERT model for encoding the question text.

10 Acknowledgments

We thank the University of Maryland's 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 thank John Kirchenbauer for his suggestions on the toolings used for experimental evaluations. We thank Ryan Rosenberg and Ophir Lifshitz for their discussions of buzzpoint data. This material is based upon work supported by the National Science Foundation under Grant No. IIS-2403436 (Boyd-Graber) and the Army Research Office under Grant Number W911NF-23-1-0013 (Gor). Any opinions, findings, views, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the official policies of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Finally, we express our gratitude to Flaticons[11] 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, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *arXiv preprint arXiv: 2311.16867*.

Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020. R2de: a nlp approach to estimating irt parameters of newly generated questions. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 412–421.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

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*.

[11] https://www.flaticon.com/

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*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv preprint arXiv: 2401.04088*.

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.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv preprint arXiv: 2402.09906*.

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.

Mark D. Reckase. 2006. 18 multidimensional item response theory. In C.R. Rao and S. Sinharay, editors, *Psychometrics*, volume 26 of *Handbook of Statistics*, pages 607–642. Elsevier.

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

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan 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.

Xiaoming Shi, Siqiao Xue, Kangrui Wang, Fan Zhou, James Zhang, Jun Zhou, Chenhao Tan, and Hongyuan Mei. 2024. Language models can improve event prediction by few-shot abductive reasoning. *Advances in Neural Information Processing Systems*, 36.

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*.

Gemini Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin

Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomas Kocisky, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, Juliette Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Yingjie Miao, Lukas Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontanon, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, Anja Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Ramasesh, Anton Briukhov, DaWoon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matthew Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara Sainath, Shane Gu, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, Sébastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Josh Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, Sébastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Kiran Vodrahalli, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravi Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Lučić, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjösund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Zeyncep Cankara, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Lora Aroyo, Zhufeng Pan, Zachary Nado, Jakub Sygnowski, Stephanie Winkler, Dian Yu, Mohammad Saleh, Loren Maggiore, Yamini Bansal, Xavier Garcia, Mehran Kazemi, Piyush Patil, Ishita Dasgupta, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Qingze Wang, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Raphaël Lopez Kaufman, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Chris Welty, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabe Barth-Maron, Craig Swanson, Dominika Rogozińska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Renshen Wang, Dave Lacey, Anastasija Ilić, Yao Zhao, Adam Iwanicki, Alejandro Lince, Alexander Chen, Christina Lyu, Carl Lebsack, Jordan Griffith, Meenu Gaba, Paramjit Sandhu, Phil Chen, Anna Koop, Ravi Rajwar, Soheil Hassas Yeganeh, Solomon Chang, Rui Zhu, Soroush Radpour, Elnaz Davoodi, Ving Ian Lei, Yang Xu, Daniel Toyama, Constant Segal, Martin Wicke, Hanzhao Lin, Anna Bulanova, Adrià Puig-

domènech Badia, Nemanja Rakićević, Pablo Sprechmann, Angelos Filos, Shaobo Hou, Víctor Campos, Nora Kassner, Devendra Sachan, Meire Fortunato, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, Tina Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anaïs White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Ying Xu, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, Çağlar Ünlü, David Reid, Zora Tung, Daniel Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Ricardo Aguilar, Mai Giménez, Jiawei Xia, Olivier Dousse, Willi Gierke, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Dan Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nick Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, DongHyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alex Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clement Farabet, Pedro Valenzuela, Quan Yuan, Ananth Agarwal, Mia Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, Andrey Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiri Simsa, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas FitzGerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel Kaed, Jing Li, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Alanna Walton, Alicia Parrish, Mark Epstein, Sara McCarthy, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv: 2403.05530*.

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 11 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 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 11: 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 CAIMIRA Setup.

In this section, we provide a detailed explanation of the learning objective for CAIMIRA and the hyperparameters used in our experiments. First, let's revise the CAIMIRA objective from Section 3:

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

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.

Here, $\mathbf{d_i}$ and $\mathbf{r_j}$ are functions of question representation $\mathbf{E}_j^q$ defined as:

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

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

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 ($\mathbf{s_i} = \mathbf{E}_i^a$), are the parameters we train for CAIMIRA over a regularized cross entropy objective.

Hyperparameters. The trainable parameters are fit using mini-batch stochastic gradient descent to minimize $\mathcal{L}_{\text{CAIMIRA}}$ (Equation 11), where λ_d and λ_s are set to $1e-5$. We use Adam optimizer (Kingma and Ba, 2014) without weight decay, and with a learning rate of 0.005, and the batch size is set to 512.

C 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.

Figure 12: Agents we use in the Context Retrievers category.

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 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

Figure 13: Agents we use in the Title Retrievers category.

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., 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. Figure 17 shows an example of the prompt used for these models.

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 18 shows an example of the prompt used for RAG models.

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

D 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.

40b+ LLMs

- cohere-command-r-plus_1shot
- falcon-40b-instruct_1shot
- falcon-40b_1shot
- llama-2-70b_1shot
- meta-llama-3-70b-instruct_1shot
- meta-llama-3-70b_1shot
- mixtral-8x7b-instruct_1shot

Inst Ctx Retriever R@10

- grit_ctx-recall@10

Inst Ctx Retriever R@3

- grit_ctx-recall@3

Inst Ctx Retriever R@1

- grit_ctx-recall@1

Base LLMs

- gpt-neo-2.7B_1shot
- opt-2.7b_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
- gemma-1.1-7b-it_1shot
- mistral-7b-inst_1shot
- opt-iml-max-30b_1shot
- phi-3-mini-3.8b_1shot

Figure 14: Agents we use in the LLMs category.

OpenAI GPT3+

- openai-gpt-3.5-turbo_1shot
- openai-gpt-4-turbo_1shot
- openai-gpt-4o_1shot

Figure 15: Agents we use in the GPT-3+ category.

RAG (Top 10)

- rag-bm25_top10-flan-ul2
- rag-bm25_wiki_top10-command-r-plus
- rag-grit_top10-flan-ul2
- rag-grit_wiki_top10-command-r-plus

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 16: Agents we use in the RAG category.

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

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

E Question Difficulty

This section enlists the full set of heatmaps of mean relevance $\mathbf{r_{j,k}}$ and *mean effective difficulty* $\mathbf{d}^{(\mathbf{e})}_{\mathbf{D},\mu_\mathbf{k}}$ of question clusters across the five latent factors (k).

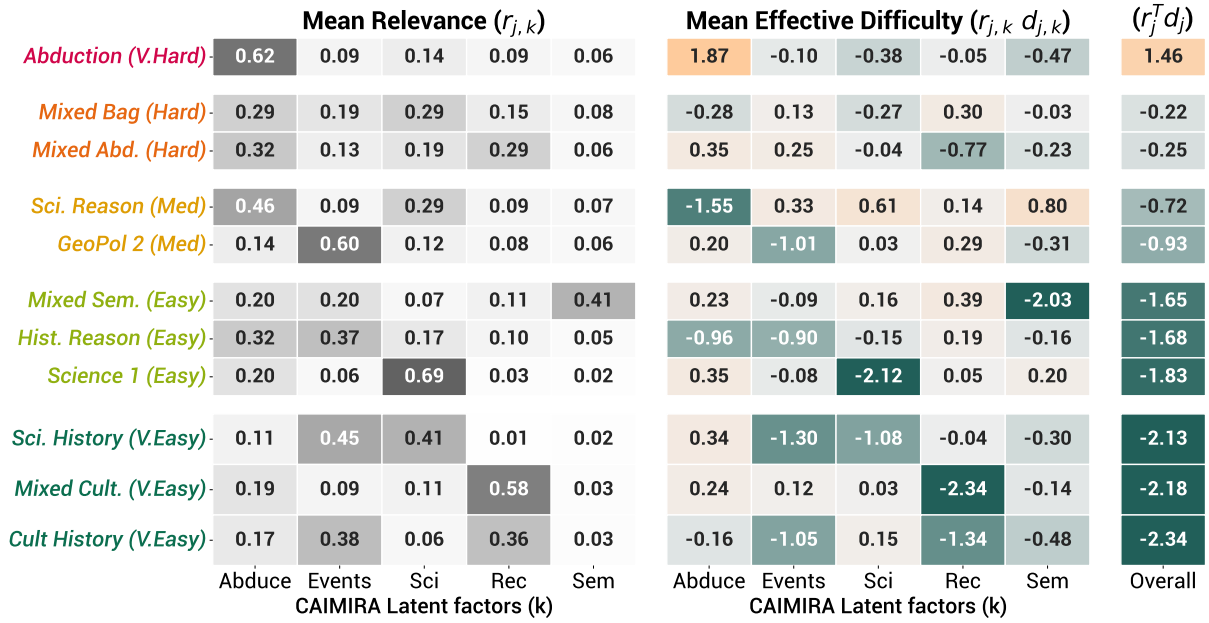

Figure 19: Heatmaps of mean relevance $\mathbf{r_{j,k}}$ and *mean effective difficulty* $\mathbf{d}^{(\mathbf{e})}_{\mathbf{D},\mu_\mathbf{k}}$ of question clusters across the five latent factors (k).

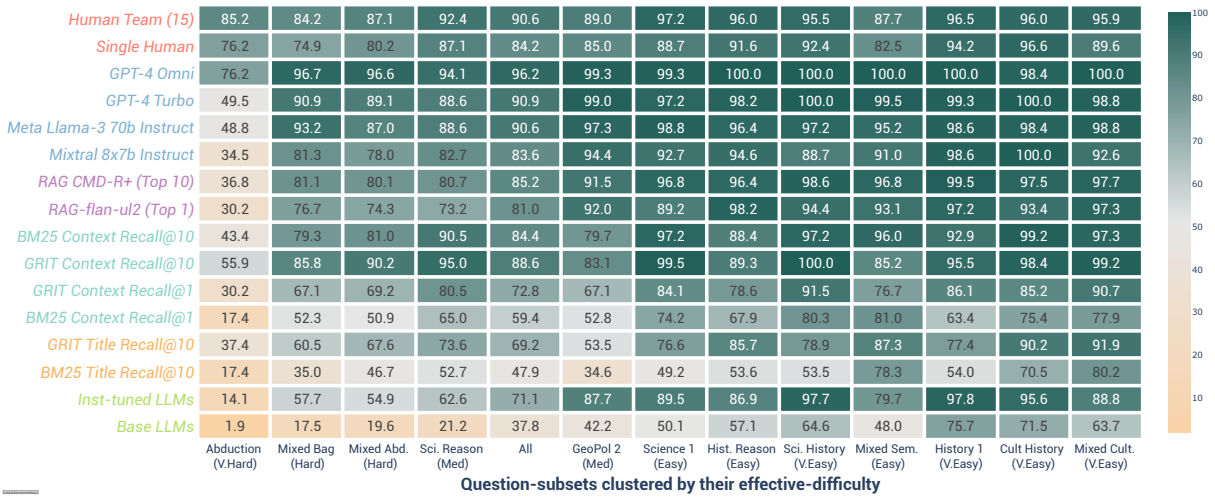

Figure 20: Full set of agent accuracies across all question clusters defined in Figure 19. We use the same color scheme as in Figure 9.

Abduction (V.Hard)

Answer: Mount Olympus
Clues: Homer claimed that this place never has storms and is bound in aether.

Answer: medians
Clues: Apollonius' Theorem can be used to find the length of this construct given the side lengths of a triangle.

Answer: The Arnolfini Marriage
Clues: Symbols in this painting include a pair of discarded clogs and a chandelier with one lit candle. In the middle of this painting, a feather duster and a beaded chain flank the artist's signature, which is above a circular mirror. A dog sits near this painting's two human figures, one of whom wears a green dress as she holds the hand of her suitor.(*)

Answer: Ramona Geraldine Quimby
Clues: This owner of a stuffed elephant named Ella Funt plays a black-nosed sheep in a Christmas play and dresses up as "the baddest witch in the world." She has a cat named Picky-Picky until it dies, and she also sees herself in an infinite mirror.

Answer: A Wrinkle in Time
Clues: Two characters in this book later appear as the main characters of Many Waters. Mrs. Whatsit, Mrs. Who, and Mrs.Which start this journey in this book.

Answer: rectangles
Clues: The uniform probability distribution takes this shape. Rotating this shape using one of its sides as an axis yields a cylinder. This shape is traced out by the x-axis, the y-axis, and the equations x equals two and y equals six.

Answer: To Kill a Mockingbird
Clues: One character in this book deliberately pours syrup all over his lunch. At one point, the main characters are taken to a church by their cook, Calpurnia.

Answer: (Alexandre) Gustave Eiffel
Clues: This man designed railway stations in Santiago, Chile and Budapest, Hungary. He was jailed after being implicated in a failed Panama Canal project, for which he designed the locks.

Answer: Lord of the Flies
Clues: In this novel, a dead parachutist is discovered by the strange introverted character Simon. Sam and Eric are the last followers of one character in this novel.

Answer: Eminem
Clues: This musician says, after declaring "now I'm gonna make you dance," "girl you know you're my world" in his song "Just Lose It."

Figure 21: Examples of questions from different clusters.

Mixed Abd. (Hard)

Answer: Justin Bieber
Clues: This singer claims "I'd wait for you forever and a day" and "your world is my world" in one song. Big Sean wonders "I don't know if this makes sense, but you're my hallelujah" in a song where this singer says he'll be your (*) platinum, silver and gold.

Answer: Neil Gaiman
Clues: This frequent collaborator of Dave McKean won both the Carnegie and Newbery Medals for a book about a crypt full of Sleer being explored by Nobody Owens.

Answer: Moby-Dick (or The Whale)
Clues: Characters in this novel include the Zoroastrian Fedallah (feh-DAH-lah), a Native American called Tashtego, and a South Sea islander named Queequeg (KWEE-KWAIG).

Answer: Samson
Clues: Before he was born, his parents learned that he was not to touch a dead body, and he was to abstain from strong drink. He was involved with a Timnite woman and a harlot before meeting the woman that would betray him.

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.

Answer: Mean
Clues: The harmonic one of n numbers in a data set is n divided by the sum of the reciprocals of the numbers. The geometric one is the nth root of the product of the numbers. The geometric one is always less than or equal to the arithmetic ("air-ith-MET-ick") one.

Answer: Alice
Clues: This character watches a lion and a unicorn fight over a crown, and although her cat Dinah will not talk to her, the Tiger Lily and the other flowers will.

Answer: Daniel
Clues: As punishment for not worshipping a golden statue, this man's friends were ordered thrown into a furnace, but they were not burned. While training to be a scribe, this man was given the Babylonian name Belteshazzar ("BEL-tuh-SHAH-zar").

Answer: magma
Clues: The three types of this material differ by their mineral and gas content; rhyolitic and andesitic types contain more silicon dioxide and are more viscous. The basaltic type is hottest, forms due to partial melting in the mantle, and flows fastest.

Answer: parallelogram
Clues: This shape names a law for adding vectors. In a namesake illusion, diagonals of two of these figures appear to be different lengths, though they are not.

Figure 22: Examples of questions from different clusters.

Mixed Bag (Hard)

Answer: prime numbers
Clues: The fundamental theorem of arithmetic states that every positive integer can be uniquely represented as a product of these numbers. Special types of these numbers are named after Fermat (“fur-MAHT”) and Mersenne (“mur-SEN”). To find these numbers, one may use the Sieve of Eratosthenes (air-uh-TOSS- then-eez”), in which one crosses off all multiples of two, then all multiples of three, and so on. For 10 points, give these numbers whose only factors are one and themselves.

Answer: gerrymandering
Clues: The Justice Department suggested using race as a basis for this practice in the 1990’s.

Answer: Secretary of State
Clues: Resignations of the President or Vice-President must be delivered to this person. Madeleine Albright was the first woman to hold this position, and one candidate for this position in the second Obama administration withdrew her candidacy due to controversy over the (*) Benghazi attacks.

Answer: Romeo and Juliet
Clues: This play’s opening brawl is started by Gregory and Samson. Later in this play, Friar John fails to deliver a letter written by Friar Lawrence.

Answer: Sagittarius
Clues: Both Globular Cluster M54, the center of this constellation’s namesake dwarf elliptical galaxy, and a possible supermassive black hole at the center of the Milky Way are found in this constellation.

Answer: photographs
Clues: An early invention used to make art works in this medium was the daguerreotype [duh-gayr-"row"-"type"]. Eadweard ["edward"] Muybridge created works in this medium which clarified the method by which horses gallop. The Steerage and Migrant Mother are specific examples of these types of art works.

Answer: sine
Clues: This function’s namesake law relates the side length to the opposite angle in any triangle.

Answer: static
Clues: This term describes a type of friction whose coefficient is usually larger than that of kinetic friction. It describes a type of equilibrium in which the net torque and net force both equal zero, resulting in a motionless object.

Answer: Peter I
Clues: This man’s reign began with the Streltsy (SHTRELT-zee) Revolt instigated by his half-sister, Sophia.

Answer: greatest common factor
Clues: Antenaresis, or Euclid’s method, can be used to find this value given any two numbers. It can be also be found by multiplying two numbers and dividing by their least common multiple.

Figure 23: Examples of questions from different clusters.

Sci. Reason (Med)

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.

Answer: tundra
Clues: Cushion plants are found in the alpine form of this biome, which is also home to marmots, pikas, and chinchillas. The point at which this biome meets taiga is known as the treeline. Flora in this biome consists of lichens (LYE-kens) and mosses. Non-alpine forms of it have little vegetation due to permafrost.

Answer: Lois Lowry
Clues: One of this writer's stories follows Annemarie Johansen as she helps her friend Ellen escape from Nazi-occupied Denmark. A sequel to this author's most well-known book follows the weaver Kira, and that book ends with Jonah and Gabe fleeing the dystopian society they live in.

Answer: calcium
Clues: Channels that carry ions made of this element are blocked by some hypertension medications.

Answer: M̈y Life Would Suck Without You¨
Clues: The protagonists of this song's music video throw magazines, clothes and an empty fishbowl out an open window. This song notes that "maybe I was stupid for telling you goodbye" regarding a boy who the singer supposes is sorry because "you're (*) standing at my door." This song's chorus notes that "you've got a piece of me and honestly" before expressing the title sentiment.

Answer: Ramona Geraldine Quimby
Clues: This owner of a stuffed elephant named Ella Funt plays a black-nosed sheep in a Christmas play and dresses up as "the baddest witch in the world." She has a cat named Picky-Picky until it dies, and she also sees herself in an infinite mirror. This best friend of Howie Kemp lives on the same street as Henry Higgins. For 10 points, name this little sister of Beezus, the main character of a series of books by Beverly Cleary.

Answer: guns
Clues: In Major Barbara, Andrew Undershaft became rich by manufacturing these objects. Both Hedda Gabler and Young Werther (VEHR-tuhr) commit suicide using these objects.

Answer: Bridge to Terabithia
Clues: This novel's protagonist wants to become the fastest runner in the fifth grade, but that plan is spoiled by the girl who moves in next door. While this book's protagonist visits the National Art Gallery with his music teacher, that girl tries to (*) swing over the creek, but the rope snaps and she dies.

Answer: Curie
Clues: Two brothers of this surname discovered piezoelectricity and a namesake point at which ferromagnetic materials become paramagnetic. One of those brothers explored the properties of the ore pitchblende with his wife. That wife later won a second Nobel Prize for her work isolating radium, and named the element polonium after her native country. For 10 points, give the last name of physicist Pierre and his wife Marie.

Answer: polls
Clues: The "straw" form of this practice is unscientific and the "push" form of this is really just a campaign tactic designed to attack an opponent in disguise.

Figure 24: Examples of questions from different clusters.

Mixed Sem. (Easy)

Answer: Richard I of England
Clues: This man was killed by a crossbow bolt while besieging the castle Charlus-Chabrol. After the departure of Philip Augustus of France, this man led the Christian armies in the Third Crusade, during which he achieved peace with Saladin. He was succeeded by his brother John. For 10 points, name this 12th-century King of England known by an epithet signifying his bravery.

Answer: Vincent (Willem) Van Gogh
Clues: While in Auvers [oh-vair], this man painted his physician holding a foxglove plant. In another painting by him, a woman pours coffee as a destitute family sits at a table for a meal. His best-known work shows Saint- Rémy [sahn-ray-mee], and this artist painted the Portrait of Dr. Gachet [gah-shay] and The Potato Eaters.

Answer: William Faulkner
Clues: In this author's first Pulitzer Prize-winning work, the Generalissimo orders the execution of Corporal Zsettslani ("SET-slah-nee"). His second Pulitzer-winning novel revolves around Lucius Priest, a resident of Yoknapatawpha ("YOCK-NAH-puh-TAH-fuh") County. This author wrote novels about Thomas Sutpen and about the death of Addie Bundren. For 10 points, name this American author of Absalom! Absalom!, As I Lay Dying, and The Sound and the Fury.

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. The Treaties of Velasco were signed following this leader's capture after the Battle of San Jacinto, and he was responsible for the deaths of Jim Bowie and Davy Crockett.

Answer: "Auld Lang Syne"
Clues: This poem's original form notes that the speaker and his addressee have "rin about the braes" and "paidl't i' the burn." The speaker of this poem written in Scottish dialect claims that they will "take a cup of kindness yet" and asks, "Should auld acquaintance be forgot, and never brought to min'?" For 10 points, name this Robert Burns poem that is often sung on New Year's Eve.

Answer: Pytor Ilyich Tchaikovsky
Clues: This musician dedicated his Symphony No. 4 in F Minor to his financial supporter Nadezhda (nah- DEZH-dah) von Meck, though they never met. His Sixth Symphony, nicknamed Pathetique (pah-theh- TEEK), premiered nine days before his death.

Answer: The Outsiders
Clues: In this novel, Bob Sheldon and Randy Adderson take part in an attack on Johnny, causing Johnny to fear for his life.

Answer: To Kill a Mockingbird
Clues: In this novel the narrator's father shoots Tim Johnson, a rabid dog. The narrator and her brother are attacked on the way home from a Halloween pageant, but are saved by Boo Radley.

Answer: Johann Sebastian Bach
Clues: Lieschen [lee-shen] is addicted to coffee in a cantata by this composer of the Notebook for Anna Magdalena. Gounod's [goo-noh's] Ave Maria is based on a prelude from this composer's Well-Tempered Clavier, and Mendelssohn revived his setting of the St. Matthew Passion.

Answer: Don Quixote de la Mancha
Clues: This character interrupts a round of storytelling by attacking a stash of wine-skins. He wears a washbasin as a helmet while calling himself the Knight of the Sorry Face. He owns the horse Rocinante (ROHsin- AHN-tay) and frequently speaks of his love for Dulcinea (dull-sin-AY-ah) to his friend Sancho Panza. For 10 points, name this self-proclaimed knight from La Mancha who fights against windmills in a book by Miguel de Cervantes.

Figure 25: Examples of questions from different clusters.

Science 1 (Easy)

Answer: Spanish
Clues: One writer in this language wrote the collection "Twenty Love Poems and a Song of Despair."

Answer: Earth
Clues: In Jainism, this object's central point is Mount Meru. In Chinese mythology, this object is the lower half of a cosmic egg split by Pangu, while in ancient Egypt the original form of this object was the primordial (*) mound.

Answer: mitochondria (" MY-toe-KON-dree-uh ")
Clues: The DNA in this organelle ("or-guh-NELL") is inherited only from the mother. The inner membrane of this organelle contains folds known as cristae ("CRISS-tay") and encloses its matrix.

Answer: coral reefs
Clues: Darwin's first paper was on the formation of this biome, whose organisms are threatened by white-band disease. Acidification removes the minerals needed for this ecosystem to grow as each new generation builds on the calcium carbonate skeletons of the previous one.

Answer: Ohio
Clues: n this state's capital, the Lane Avenue Bridge crosses the Olentangy River. Another of its cities contains historic Italian architecture in its Over-the-Rhine neighborhood, while another city, at the mouth of the Cuyahoga River, contains Case Western Reserve University. Much of its northern border is at Lake (*) Erie, and it is separated from Kentucky by its namesake river. For 10 points, name this state containing Cincinnati, Cleveland, and Columbus.

Answer: Chlorine or Cl
Clues: Stomach acid consists mainly of a compound of hydrogen and this element. It is the second-lightest halogen, after fluorine, and at room temperature is a yellow-green gas. Compounds with it, carbon, hydrogen, and fluorine deplete the ozone layer and are called (*) CFCs. It is used in bleach as well as to disinfect swimming pools, and forms table salt along with sodium. For 10 points, name this element, number 17, symbolized Cl.

Answer: electron
Clues: This particle was discovered by J.J. Thomson, and its exact charge was discovered in the Millikan oil drop experiment. According to the Pauli Exclusion Principle, two of these particles cannot exist in the same quantum state.

Answer: matter
Clues: The density parameter for the non-relativistic form of this falls off with the cube of the scale factor. This substance dominated the universe from approximately 75,000 years after the Big-Bang until about 4 billion years ago.

Answer: violin
Clues: The Rhapsody on a Theme of Paganini was written from twenty-four caprices originally written for this instrument. Vivaldi's The Four Seasons is a set of concerti ("con-CHAIR-tee") written for this instrument.

Answer: glaciers
Clues: These objects contain the zone of plastic flow and the zone of brittle flow. They are formed by compressing firn, and parts of them break off by calving. Till is soil left behind by these objects, which also push material to form moraines.

Figure 26: Examples of questions from different clusters.

Hist. Reason (Easy)

Answer: Scooby-Doo
Clues: Big Bob Oakley was the first person on this show to say "I'd have gotten away with it too, if it weren't for those kids," and one show in this series introduced a character named Scrappy. In 2002, a film of the same name starred Freddie Prinze, Jr. as Freddy and Sarah Michelle Gellar as Daphne. For 10 points, name this cartoon franchise, named for a cowardly Great Dane.

Answer: Steve Jobs
Clues: This man, along with Edwin Catmull, was credited as an executive producer of the original Toy Story movie, produced by Pixar Animation, which he renamed after purchasing it from George Lucas in 1986. From 2000 to 2011, he served as CEO of the computer company he co-founded with Steve Wozniak.

Answer: Neptune
Clues: A triangular patch of clouds that circulates this planet quickly is known as The Scooter. Its atmosphere contains the fastest winds in the solar system. Its existence was predicted by Alexis Bouvard, and it was discovered by Johann Galle. It often contains the Great Dark Spot. Its largest moon, which has a retrograde orbit, is Triton. For 10 points, name this gas giant, the farthest from the Sun in the solar system.

Answer: Orion
Clues: This constellation contains the Trapezium Cluster and is the site of a late-October meteor shower.

Answer: Niccolo Machiavelli
Clues: Although he is not Sun Tzu, this man wrote a version of The Art of War. He wrote a critique of Roman history in his Discourses on Livy.

Answer: prime numbers
Clues: The fundamental theorem of arithmetic states that every positive integer can be uniquely represented as a product of these numbers.

Answer: The New York Times
Clues: This newspaper was sued by Alabama public safety officer Louis B. Sullivan. Its long-time publisher, Arthur Ochs Sulzberger, died in 2012.

Answer: Uncle Tom's Cabin
Clues: In this novel, shelter is provided by the Halliday and Bird families. At the beginning of this novel, the Shelby family sells their property to the St. Clare family. At the end of this novel, George and Eliza Harris escape north. The husband of Aunt Chloe is killed by Simon Legree in, for 10 points, what American novel, depicting the life of slaves, written by Harriet Beecher Stowe?

Answer: Harry Mason Reid
Clues: This man almost lost his Senate seat in the 1998, surviving a challenge from future colleague John Ensign, and he is expected to have a tough re-election in 2010 against Sue Lowden or Danny Tarkanian. He commented that Barack Obama was "light-skinned" and "spoke with no Negro dialect, unless he wanted one." For 10 points, name this senior Senator from Nevada, the current Senate Majority Leader.

Answer: Pangaea
Clues: One piece of evidence that supports its existence is that the Caledonian mountains of Northern Europe are a continuation of the Appalachian Mountains. This entity broke up into Laurasia and Gondwanaland ("gon-DWON-uh-land").

Figure 27: Examples of questions from different clusters.

History 1 (V.Easy)

Answer: Puerto Rico
Clues: The independence of this commonwealth has been sought by Rubén Berríos, while an opposite approach has been pushed by its New Progressive Party under Pedro Pierluisi. In 2012, this commonwealth elected Alejandro García Padilla as governor and voted in a referendum to end its territorial status. (*) For 10 points, name this Caribbean Island, a United States territory that may someday become the 51st state.

Answer: Philadelphia, Pennsylvania
Clues: In this city, Wissahickon Creek goes through Fairmount Park. This city can be entered by crossing the Delaware River on the Betsy Ross Bridge. One of its buildings, where the Second Continental Congress adopted the (*) Declaration of Independence, is Independence Hall. The Liberty Bell is found in, for 10 points, what city in Pennsylvania?

Answer: Yellowstone National Park
Clues: The last wild herd of bison in the United States was located in this park, where today they are hunted by grizzly bears and wolves reintroduced in the 1990s.

Answer: Leo Tolstoy
Clues: One work by this author, about a man who injures himself while hanging curtains, is The Death of Ivan Ilyich. One of his novels has a relationship between Levin and Kitty, while the title character has an affair with Count Vronsky and eventually commits suicide by jumping in front of a (*) train. For 10 points, name this author who wrote about the French invasion of Russia in War and Peace in addition to writing Anna Karenina.

Answer: Federal Republic of Germany
Clues: One leader of this country forcibly annexed the Sudetenland ("soo-DAY-ten-land"). During a movement to reunite this country, the leader of one half operated under the policy of ostpolitik ("OST-pol- it-ick"). Following World War I, the Weimar ("VIE-mar") Republic was established in this nation.

Answer: Thomas Jefferson
Clues: This politician responded to Francois Barbe-Marbois in his Notes on the State of Virginia. This man founded the University of Virginia and designed the mansion of Monticello..

Answer: Mexico
Clues: In 1822, the House of Iturbide ("EE-tur-BEE-day") assumed control of this nation for one year. This nation was ruled by an Austrian emperor installed by Napoleon III, Maximilian, although he was overthrown by Benito Juarez ("WAHR-ezz"). The Gadsden Purchase bought land from this country, whose victory at Puebla ("PWAY-bluh") is celebrated as Cinco de Mayo. For 10 points, identify this nation that once owned California and Texas.

Answer: Ronald (Wilson) Reagan
Clues: This man used powers granted by the Taft-Hartley Act during a confrontation with air traffic controllers, and his Defense Secretary resigned after violations of the Boland Amendment were revealed. Before those events during his presidency, he served as Governor of California from 1967 until 1975. Prior to entering politics, this man was a famous (*) Hollywood actor. For 10 points, name this Republican president from 1981 to 1989.

Answer: Isaac Asimov
Clues: This author wrote a story in which the inhabitants of Lagash experience darkness for the first time. Along with "Nightfall," this author wrote a series of novels featuring the investigative interactions of Elijah Baley and R. Daneel Olivaw. Hari Selden invents the science of psychohistory in this author's novel (*) Foundation. For 10 points, name this Russian-American science fiction writer who depicted the Three Laws of Robotics in his collection, I, Robot.

Answer: Julius Caesar
Clues: This man fought against Ariovistus ("air-ee-oh-VIS-tuss"), a German leader, and Vercingetorix ("ver-KING-uh-TOR-ix"), a chieftain of the Arverni ("ar-VEHR-nee") whose defeat is described in this man's book, Commentaries on the Gallic Wars. He led his troops across the Rubicon to start a civil war with Pompey, one of his partners in the First Triumvirate. For 10 points, name this Roman leader who was assassinated by Brutus on the Ides of March.

Figure 28: Examples of questions from different clusters.

Mixed Cult. (V.Easy)

Answer: The Nutcracker
Clues: This work opens with the title item given as a gift by Drosselmeyer; it is later broken by Fritz. Spanish, Arabian, and Chinese dances in this ballet are said to represent different substances such as chocolate, coffee, and tea. The Waltz of the Snowflakes and Dance of the (*) Sugarplum Fairy appear in, for 10 points, what Peter Tchaikovsky ballet about Clara's Christmas gift coming to life?

Answer: King Arthur
Clues: A popular novel about this figure is T.H. White's The Once and Future King. In the Annales Cambriae (ah-NAH-less CAM-bree-ay), this figure was mortally wounded at the Battle of Camlann during a fight with his son Mordred.

Answer: Thebes
Clues: This city was founded by Cadmus after following a cow until it sat. This city was besieged by the Sphinx, as all travelers who entered it were forced to either solve its riddle or be eaten. To avenge the sleight done to him by Eteocles("et-TEE-oh-clees"), Polyneices ("polly-NYE-kees") led a group of seven warriors against this city.

Answer: WikiLeaks
Clues: A PowerPoint presentation released by this organization details how Bank of America plans to attack it. One portion of this organization is run by the Sunshine Press. In November 2010, a Fox News host called it a "terrorist organization" after it published U.S. State Department diplomatic cables.

Answer: Isaac Newton
Clues: In this scientist's book Opticks, he discussed his experiments with the dispersion of light, including breaking white light into its constituent colors using a prism. One law named for him describes "universal (*) gravitation"; another states that the net force on an object is its mass times its acceleration, while a third states that for every action there is an equal and opposite reaction. For 10 points, name this English scientist who formulated three laws of motion.

Answer: Girl Scout Cookies
Clues: A group from Muskogee, Oklahoma is believed to be the first to produce and sell these items popularly sold as a fundraiser for an organization founded by Juliette Gordon Low in 1912.

Answer: Odysseus
Clues: This man's dog Argus dies atop a refuse heap. He reveals himself to a foot-washing maid, Eurycleia ("your-ee-CLAY-uh"). The Laestrygones ("LAY-strih-GOAN-ees") destroy many ships belonging to his fleet, and he also visits the land of the lotos ("lotus") -eaters. He kills his wife's suitors with the help of his son, Telemachus ("TELL-uh-MOCK-us"), then reunites with that wife, Penelope. For 10 points, an epic by Homer describes what man's twenty-year quest to get home after the Trojan War?

Answer: Alice
Clues: This character watches a lion and a unicorn fight over a crown, and although her cat Dinah will not talk to her, the Tiger Lily and the other flowers will. She shrinks after drinking a potion labeled "Drink Me," and attends a tea party with a sleepy Dormouse, a March Hare, and a Mad Hatter.

Answer: Trojan War
Clues: Neoptolemus killed King Priam in the final stages of this event, after which Aeneas fled with his son. This event began after the Judgement of Paris and (*) Helen's abduction from King Menelaus of Sparta. After nine years, it finally ended after Greek soldiers got past enemy gates while hiding in a giant wooden horse. For 10 points, name this conflict in Greek mythology that featured warriors like Hector and Achilles.

Answer: Noah
Clues: Seven laws that apply to non-Jews are named for this figure, whose nakedness was uncovered by one of his sons. An agreement this figure made with God is symbolized by the rainbow. He was the son of Lamekh (LAH-meck) and had three sons, Japheth (JAY-feth), Ham, and Shem. To confirm that one of his jobs was complete, he sent a dove to check for dry land. For 10 points, identify this Biblical character who took two animals of each kind in his ark.

Figure 29: Examples of questions from different clusters.

Sci. History (V.Easy)

Answer: Andes Mountains
Clues: This mountain range includes the Vilcabamba (“VEEL-cuh-BOM-buh”) sub-range and contains a plateau called the altiplano (“ALL-tee-PLAN-oh”).

Answer: London
Clues: Hampstead Heath and Kensington Gardens are parks in this city which is served by the "Jubilee Line," "Piccadilly Line," and "Victoria Line" of its subway system, the Underground. A Norman castle built by William the Conqueror is this city’s "Tower."

Answer: Amazon River
Clues: The island of Marajo (mah-RAH-hoh) is located at the mouth of this river which was named by Spanish conquistador Francisco de Orellana (day OH-ray-YAH-nah) for the warrior women of Greek mythology.

Answer: Panama Canal
Clues: Lake Gatun (“GAH-tune”) is part of this waterway, whose construction was made possible by the Hay-Bunau-Varilla (“HAY boo-NOW vah-REE-uh”) Treaty and the secession of a province from Colombia. A 1977 agreement between Omar Torrijos (“torr-EE-hos”) and Jimmy Carter resulted in the return of the special zone associated with it.

Answer: Antarctica
Clues: This geographical feature has its lowest point at Bentley Trench. A lake here lies under Vostok Station. Mt. Erebus is found on Ross Island off itscoast, between Marie Byrd and Victoria lands. The Sentinel Range of the Ellsworth Mountains contains its highest peak, Vinson Massif, located on the Ronne (*) Ice Shelf.

Answer: Saturn
Clues: Great White Spots are frequent storms on this planet. Its moons include Iapetus, Rhea, Enceladus, and the only known one to have an atmosphere. This planet is less dense than water. The Cassini Division is located in its extensive ring system. For 10 points, name this second largest planet in the solar system, the sixth from the Sun.

Answer: New York City
Clues: A museum branch located in this city’s Fort Tryon Park containing medieval art is known as The Cloisters. One of its straits, which includes Roosevelt Island and Rikers Island, is the East River.

Answer: Panama Canal
Clues: Lake Gatun (“GAH-tune”) is part of this waterway, whose construction was made possible by the Hay-Bunau-Varilla (“HAY boo-NOW vah-REE-uh”) Treaty and the secession of a province from Colombia.

Answer: Vienna, Austria
Clues: This city contains the neo-gothic Votive Church, and its Karlskirche (KARLS-keer-kuh) is the largest Baroque Cathedral north of the Alps. It is the capital of a country with such states as Burgenland, Tyrol, and Styria. This city’s Ring Boulevard was ordered to be restructured by Franz Joseph I, and it lies on the Danube just upriver from Bratislava, the capital of Slovakia.

Answer: Orion
Clues: This constellation contains the Trapezium Cluster and is the site of a late-October meteor shower. One of its stars, formerly known as the Amazon Star, is Bellatrix, and its brightest stars are Betelgeuse and Rigel. Its namesake nebula joins with Hatysa and other stars to form its sword, while Alnitak, Alnilam, and Mintaka form its belt.

Figure 30: Examples of questions from different clusters.

Cult History (V.Easy)

Answer: Michelangelo di Lodovico Buonarroti Simoni
Clues: This artist's statues of a dying slave and a horned Moses were to adorn the tomb of Julius II. His only signed work is one in which Mary holds the dead body of Jesus, entitled Pietá ("pee-AY-tuh"). One of his works depicts a nude giant killer holding a sling.

Answer: Charles Dickens
Clues: This author wrote about the eviction of Nell Trent and her grandfather from The Old Curiosity Shop. In another work by this author, Abel Magwitch raises a fortune for the orphan Pip, who loves Estella. He also wrote about Sydney Carton sacrificing himself to save Charles Darnay in a work set in London and Paris.

Answer: Oklahoma
Clues: This modern state's panhandle was crossed by the Cimarron Cutoff, a branch of the Santa Fe Trail. A city in this state is called "Broken Arrow" because it was settled by Creek people, while part of this state was known as the "Indian Territory." White settlers who anticipated an 1889 decision to open its lands to homesteaders gave this state its nickname: the Sooner State. For 10 points, Tulsa is located in what state between Texas and Kansas?

Answer: Blessed Virgin Mary
Clues: In the Gospel of James, this Biblical figure is described as the child of Anna and Joachim. At the First Council of Ephesus, this figure was given the epithet Theotokos, or "God-Bearer." Martin Luther described this person as "the highest woman." This woman is held to be free from original sin under the doctrine of Immaculate Conception. For 10 points, name this mother of Jesus of Nazareth.

Answer: Frankenstein, or the Modern Prometheus
Clues: The protagonist of this work returns home from the University of Ingolstadt to find that Justine Moritz has been accused of his brother William's murder. The title character, whom Robert Walton discovers in the Arctic in a frame story, had earlier married Elizabeth Lavenza, who was killed on their wedding night.

Answer: Paul Ryan
Clues: This politician claimed that he went into politics because of Ayn Rand and made Atlas Shrugged required reading for his staff, but he later said he rejected Rand's atheism. He is the current chair of the House Budget Committee, and one of his budget proposals was titled (*) "The Path to Prosperity." For 10 points what Wisconsin Republican was Mitt Romney's Vice Presidential nominee in the 2012 election?

Answer: cerebrum
Clues: This structure is divided into Brodmann areas, and develops from the telencephalon ("TEAL"-en-SEFF-ah-"lawn"). The corpus callosum ("CORE"-puss kuh-LOE-sum) connects the two hemispheres of this structure, which is divided into temporal, parietal, occipital, and frontal lobes.

Answer: Michelangelo di Lodovico Buonarroti Simoni
Clues: This artist's statues of a dying slave and a horned Moses were to adorn the tomb of Julius II.

Answer: John Quincy Adams
Clues: This person negotiated a treaty that ceded Florida to the United States with Luis de Onis (loo-EES day oh-"NIECE") while serving as James Monroe's Secretary of State. This man agreed to name Henry Clay Secretary of State in order to break a deadlock in the House of Representatives; that decision was the first "corrupt bargain."

Answer: Sarah Palin
Clues: This person's visit to Fort Bragg caused a stir when the press was denied entry to a book tour for Going Rogue. This person resigned from the position of Governor of the state closest to Russia shortly after a campaign loss in the most recent general election. Tina Fey did a notable impression of, for 10 points, what unsuccessful vice presidential candidate who ran alongside John McCain in 2008?

Figure 31: Examples of questions from different clusters.