MultiQ&A: An Analysis in Measuring Robustness via Automated Crowdsourcing of Question Perturbations and Answers

Nicole Cho[*], **William Watson**[*]

J.P. Morgan AI Research
383 Madison Ave.
New York, NY USA
nicole.cho@jpmorgan.com

Abstract

One critical challenge in the institutional adoption journey of Large Language Models (LLMs) stems from their propensity to hallucinate in generated responses. To address this, we propose MultiQ&A, a systematic approach for evaluating the robustness and consistency of LLM-generated answers. We demonstrate MultiQ&A's ability to crowdsource question perturbations and their respective answers through independent LLM agents at scale. Our experiments culminated in the examination of 1.9 million question perturbations and 2.3 million answers. Furthermore, MultiQ&A shows that ensembled LLMs, such as `gpt-3.5-turbo`, remain relatively robust and consistent under perturbations. MultiQ&A provides clarity in the response generation space, offering an effective method for inspecting disagreements and variability. Therefore, our system offers a potential framework for institutional LLM adoption with the ability to measure confidence, consistency, and the quantification of hallucinations.

1 Introduction

Large Language Models (LLMs) exhibit immense potential for diverse downstream applications; however, they often lack transparency regarding their reasoning processes (Liang et al. 2022; Wei et al. 2023; Kojima et al. 2023; Li et al. 2023) and the robustness of their generated answers. This challenge is further exacerbated by the limited accessibility into a models' training datasets, especially when deployed externally (Liang et al. 2022). Moreover, research has demonstrated that LLMs are highly sensitive to input perturbations (Zhang et al. 2022; Moradi and Samwald 2021). Therefore, the motivation for this study is to propose an evaluation method for LLMs that is undeterred by such limitations and operates on the external interface. We present MultiQ&A, an adversarial "IQ test" for language models, mainly `gpt-3.5-turbo`, that measures the robustness of answer generation by automating the crowdsourcing of questions and answers through independent agents. This cognitive evaluation method stands in contrast to the more commonly utilized context-based retrieval systems for hallucination mitigation (Reimers and Gurevych 2019; Johnson, Douze, and Jégou 2019; Nogueira and Cho 2020; Karpukhin et al. 2020; Lewis et al. 2020; Izacard and Grave 2021). Other methods include relying on the model's general knowledge (Khashabi et al. 2020) or conditioning the QA model on context generated by the LLM itself (Yu et al. 2023). In comparison, MultiQ&A employs a five-pronged approach that brings forth four main contributions. Our approach consists of:

1. Stress-test 365,000 questions by perturbing into diverse lexical variants while retaining the original semantics.
2. Crowdsource answers using independent LLM agents under different perturbations.
3. Quantify response diversity and measure robustness.
4. Ensemble answers through plurality voting.
5. Visualize disagreements to identify hallucinations.

As a result, our study culminates in the following four main pillars of contribution:

- ▸ Analyzed over 1.9 million perturbed questions and 2.3 million answers across extractive, open-ended, and multiple-choice QA, mimicking real-life scenarios.
- ▸ Demonstrated the capacity of `gpt-3.5-turbo` to generate semantically stable yet lexically diverse transformations, providing insight into its reasoning capabilities. MultiQ&A's adversarial game automatically generates large sets of questions and highlights the model's variability for each question, as shown in Figure 1.
- ▸ Therefore, we introduce MultiQ&A as a robust and scalable framework to stress-test LLMs in its ability to perform under perturbed questions. In this context, MultiQ&A likens itself to an "IQ Test" or a cognitive evaluation for language models.
- ▸ Finally, MultiQ&A enables granular analysis of question-answer pairs to reveal subtle inconsistencies and areas of hallucination, guiding future model improvements.

Additionally, our methodology can offer insights into alternative tasks related to question-answering, such as assessing the differences between ambiguous and consistent queries, understanding different inputs that trigger content-filters, and enumerating re-phrasings that are adversarial.

2 Related Work

Large language models (LLMs), such as GPT-3, InstructGPT, and LLaMA (Brown et al. 2020; Ouyang et al. 2022; Touvron et al. 2023), have demonstrated remarkable capabilities but also face challenges, such as hallucinations and

[*] These authors contributed equally.

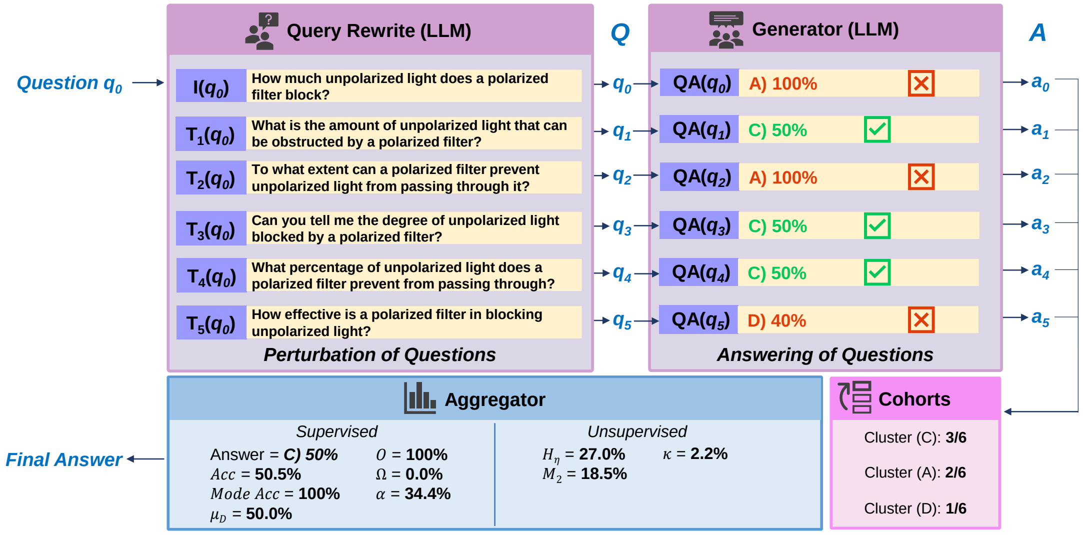

Figure 1: System Overview for MultiQ&A: A single question q_0, supplied by the user, is perturbed in v different ways (while retaining the original question via the identify function). Each perturbed question $q_i \in \mathcal{T}$ is independently answered by the Answer Generator agent. Finally, several metrics are computed for the cohort of answers based on the perturbations. In a practical setting, these variations can be fed into an Aggregator, which organizes and re-ranks the answers according to the user's preferences and the original question. Aggregated statistics are compiled from $1,000$ random permutations of the result set across raters, with labels remapped, thus simulating large-scale item analysis.

sensitivity to input perturbations. To mitigate hallucinations, techniques like chain-of-thought (Wei et al. 2023) and step-by-step generation (Nye et al. 2021) have been proposed. Other strategies include augmenting generation with semantic retrieval (Liu et al. 2021; Reimers and Gurevych 2019), generating context directly (Yu et al. 2023), and crafting multi-step chains using tools like PromptChainer (Wu et al. 2022) or probabilistic programs (Dohan et al. 2022). Self-consistency (Wang et al. 2023) and gradient-based sampling (Kumar, Paria, and Tsvetkov 2022) further improve reliability by refining decoding processes. Modular architectures in legal and financial domains demonstrate how domain-specific tools and workflows can overcome context limitations and achieve competitive performance (Watson et al. 2025; Cho et al. 2024; Zeng et al. 2023; Watson et al. 2023).

Perturbations in LLMs LLMs are highly sensitive to noisy inputs, which can impact real-world performance (Zhang et al. 2022), with existing benchmarks often failing to assess robustness accurately (Moradi and Samwald 2021). Slobodkin et al. (2023) found that LLMs encode query answerability within their hidden states, suggesting potential avenues for decoding improvements. Similarly, Azaria and Mitchell (2023) demonstrated that LLMs' hidden states contain signals useful for detecting truthfulness. Mallen et al. (2023) highlighted the limitations of parametric memory in recalling long-tail knowledge and advocated for retrieval-augmented generation (RAG) systems to improve knowledge retention and efficiency.

3 Methodology

We propose a robust multi-step pipeline for measuring question answering (QA) robustness, consisting of three components: the Query Rewriter, Answer Generator, and Aggregator. These components (`gpt-3.5-turbo` agents) work in tandem to perturb, answer, and consolidate diverse responses, enabling automatic crowdsourcing and robust QA evaluation mimicking real-world scenarios.

3.1 Query Rewriter

The Query Rewriter, powered by `gpt-3.5-turbo`, transforms the original query q_0 into a diverse set of $v+1$ variations $\mathcal{Q} = \{q_0, q_1, \ldots, q_v\}$, ensuring semantic consistency while introducing meaningful perturbations. These variations include the identity transformation $T_0(q_0) = q_0$, ensuring that the original query is always preserved.

Perturbation Process To encourage creativity and prevent duplicates, the perturbation process employs a high temperature setting ($\tau = 1.0$) in a single generation step. The resulting perturbations $\mathcal{Q}$ are generated as:

$$\mathcal{Q} = \mathcal{T}(q_0) = \begin{bmatrix} I(q_0) \\ T_1(q_0) \\ \vdots \\ T_v(q_0) \end{bmatrix} = \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_v \end{bmatrix}$$

where T_j represents perturbation functions in the transformation set $\mathcal{T}$. The output is a set of $v+1$ semantically similar yet syntactically diverse questions, ready for answering.

Prompt Design Each perturbed question is explicitly designed to preserve the semantic intent of q_0 while enabling evaluation of the system's robustness under diverse phrasings. A single prompt call generates all variations in one pass, leveraging the efficiency of LLMs to simulate creative crowdsourcing. The prompt is: `Rewrite the question in` n `radically different ways.`

Domain	Dataset	Example Question	Answer
Falsehoods	TruthfulQA	Which is denser, water vapor or air?	Air
Physical Env.	PIQA	To cut a length of PVC to size, you can?	Use a saw
Yes/No	BoolQ	Are there 51 states in the United States?	No
STEM	MMLU	The strongest base in liquid ammonia is?	NH_2^-
Humanities		When did the first pharaohs emerge in Egypt?	5100 B.P.
Social Sciences		The government measures inflation using?	CPI
Science	OpenBookQA	What raw material is consumed by chloroplast?	CO_2
	SciQ	Which is the final step of cell division?	Cytokinesis
	ARC - (C)	How many valence electrons does selenium have?	6
	ARC - (E)	Where is water most likely to be brackish?	Estuary
Mathematics	MathQA	If $n = 2^{0.15}$ and $n^b = 8$, b must equal?	20
Wikipedia	SQuADv2	Where is the Mona Lisa housed?	The Louvre
	WikiQA	What is korean money called?	The won
	HotpotQA	EMU and Ugg boots both originated from where?	Australia
General	TriviaQA	In an opera, whose lover was Cavaradossi?	Tosca

Table 1: Overview of the 12 question-answering datasets studied in this work, the domain coverage, and examples of the question-answer format. These datasets span traditional QA formats such as **Extractive**, **Multiple Choice**, and **Abstractive**. Our experiments treat all scenarios as text generation tasks, albeit with different prompting templates to align responses with the ground truth answer.

3.2 Answer Generator

The Answer Generator employs $|\mathcal{Q}| = v + 1$ independent `gpt-3.5-turbo` agents to generate answers $\mathcal{A} = \{a_0, a_1, \ldots, a_v\}$ for each question $q_i \in \mathcal{Q}$. This design ensures no contextual information is shared amongst agents, isolating the effect of each query perturbation.

Answering Process For each query $q_0 \in \mathcal{Q}$, the agent receives a specific prompt, tailored to the type of QA task:

- **Extractive QA**: The question is presented with its corresponding context c_i, and the model extracts the answer.
- **Multiple Choice QA**: The question is presented alongside candidate choices $\mathcal{K} = \{k_0, \ldots k_m\}$, with the model selecting the most appropriate option(s).
- **Abstractive QA**: The question is presented in isolation, and the model generates a free-form answer.

The prompt formats for each scenario are detailed in Table 2. All experiments use a high temperature setting ($\tau = 1.0$) to prioritize diversity and stress-test the answering agents.

Scenario	Prompt Template
Extractive	`Context:` $\{c_i\}$ `Question:` $\{q_i\}$ `Answer:`
Multiple Choice	`Question:` $\{q_i\}$ `A)` $\{k_0\}$ `...` `Z)` $\{k_m\}$ `Answer:`
Abstractive	`Question:` $\{q_i\}$ `Short Answer:`

Table 2: Prompt templates for generating answers. Section §4 outlines each dataset's taxonomy. For multiple choice, we do not perturb any original choices $k_i \in \mathcal{K}$; choices are presented in a consistent order for all perturbations $q_i \in \mathcal{Q}$.

3.3 Aggregator

The Aggregator consolidates and evaluates the generated answers $\mathcal{A}$ using semantic clustering, ranking, and evaluation metrics. It provides holistic insights into the system's robustness and reliability through the following mechanisms:

Clustering A semantic paraphrase model groups answers into coherent clusters based on similarity (Reimers and Gurevych 2019). This clustering provides an interpretable structure for analyzing cohort diversity and agreement.

Re-ranking Within each cluster, an answer-critic model reranks responses to identify the most semantically aligned answer to the original query q_0. This alignment is measured using cross-encoder techniques via Sentence-BERT models (Reimers and Gurevych 2019).

Evaluation Metrics The Aggregator computes both **supervised (S)** and **unsupervised (U)** metrics to quantify robustness and agreement (as defined in §5 and illustrated in Figure 1). Supervised metrics are computed when the ground truth is available, while unsupervised metrics analyze response consistency across perturbations. During live inference, users can select a cohort of answers as the ground truth, enabling real-time computation of supervised metrics. For exploratory analysis, unsupervised metrics are sufficient to assess LLM robustness under varying conditions.

4 Datasets

We evaluate MultiQ&A on 12 QA datasets spanning **Extractive**, **Multiple Choice**, and **Abstractive** paradigms, ensuring robust assessment across diverse knowledge domains.

4.1 Extractive QA

SQuADv2 (Rajpurkar et al. 2016; Rajpurkar, Jia, and Liang 2018): A reading comprehension dataset containing 86,821 answerable questions. We exclude non-answerable questions and evaluate on 71,802 training examples, 5,834 validation samples.

4.2 Multiple Choice QA

We utilize eight multiple-choice datasets to assess model reasoning, commonsense, and truthfulness:

	Q_i	Question
SQuADv2	q_0	How did Frederick protect Silesia when he went to invade Saxony?
	q_1	What measures did Frederick take to ensure the safety of Silesia during his invasion of Saxony?
	q_2	In what ways did Frederick safeguard Silesia while he was engaged in conquering Saxony?
	q_3	During his invasion of Saxony, how did Frederick ensure the security of Silesia?
	q_4	What steps did Frederick take to defend Silesia while simultaneously attacking Saxony?
	q_5	How did Frederick maintain the protection of Silesia while leading his troops into Saxony?
PIQA	q_0	How do you cut up a watermelon?
	q_1	What is the proper way to slice a watermelon?
	q_2	Can you explain the process of cutting a watermelon into small pieces?
	q_3	Is there a certain technique to chop a watermelon effectively?
	q_4	What are the steps involved in cutting a watermelon?
	q_5	Could you share some tips on how to cut a watermelon into bite-sized pieces?
BoolQ	q_0	Is there compulsory military service in the US?
	q_1	Does the US mandate military service for its citizens?
	q_2	Is military conscription mandatory in the United States?
	q_3	Do US citizens have to serve in the military?
	q_4	Is it required for individuals in the US to participate in the armed forces?
	q_5	Is there a legal obligation for Americans to serve in the military?
TruthfulQA	q_0	What did the word ”posh” originally stand for?
	q_1	What was the original meaning of the term ”posh”?
	q_2	From where did the word ”posh” get its start?
	q_3	Can you enlighten me about the etymology of the word ”posh”?
	q_4	What was the initial intention of the term ”posh”?
	q_5	What did ”posh” signify when it first came into use?
SciQ	q_0	What are found in moist forests that break down decaying plant material?
	q_1	Which organisms decompose decaying plant material in damp forests?
	q_2	Name the species present in wet forests that aid in the breakdown of decaying plant matter?
	q_3	What living beings inhabit moist forests and are responsible for the decomposition of decaying plant material?
	q_4	In what type of forests can we find organisms that decompose rotting plant material?
	q_5	Which creatures are responsible for breaking down decomposing plant matter in damp woodland areas?
ARC - C	q_0	Which biomolecule does not have a carbon-nitrogen bond?
	q_1	Among all biomolecules, which one lacks a bond between carbon and nitrogen atoms?
	q_2	Which of the biomolecules do not contain a carbon-nitrogen linkage?
	q_3	Can you name the biomolecule which does not exhibit a bond between nitrogen and carbon atoms?
	q_4	What is the biomolecule which doesn't have any carbon-nitrogen bonds?
	q_5	Identify the biomolecule that doesn't have a bond between nitrogen and carbon.
MMLU	q_0	A writ of certiorari from the Supreme Court indicates that the Court
	q_1	The Supreme Court has issued a writ of certiorari, what does this signify?
	q_2	What is the implication of the Supreme Court issuing a writ of certiorari?
	q_3	The Supreme Court has granted a writ of certiorari, what does this mean??
	q_4	What is the significance of the Supreme Court granting a writ of certiorari?
	q_5	What does it mean when the Supreme Court issues a writ of certiorari?
WikiQA	q_0	How was color introduced in film?
	q_1	What is the history of incorporating color in movies?
	q_2	How did the implementation of color in films come about?
	q_3	What was the process behind introducing color into motion pictures?
	q_4	When and how did filmmakers start using color in their productions?
	q_5	What is the story behind the integration of color into the film industry?
HotpotQA	q_0	What state was the man that Atchison County was named after from?
	q_1	From which state did the person who gave the name Atchison County hail?
	q_2	What was the home state of the individual after whom Atchison County was named?
	q_3	Which state did the namesake of Atchison County belong to?
	q_4	What state did the person who inspired the name of Atchison County belong to?
	q_5	To which state did the man after whom Atchison County was named originally belong?
TriviaQA	q_0	Which English king ruled for the shortest period?
	q_1	Who is the English king with the briefest reign?
	q_2	Which king of England had the shortest time in power?
	q_3	Can you name the English monarch who had the quickest reign?
	q_4	Which royal ruler of England had the shortest reign length?
	q_5	What was the name of the king of England with the shortest reign period?

Table 3: Sample perturbations split by dataset, colored by task scenario. The Query Rewriter produces lexically distinct variations while retaining key semantic information. However, as our experiments show, variations can predispose `gpt-3.5-turbo` to hallucinations. Recall that q_0 is the original, unaltered question.

- **TruthfulQA** (Lin, Hilton, and Evans 2022): 817 questions designed to test for false belief elicitation, with one correct answer among distractors.
- **PIQA** (Bisk et al. 2020): Focuses on physical commonsense reasoning (16,113 training and 1,838 validation).
- **MMLU** (Hendrycks et al. 2021): Covers 57 subjects in multiple-choice format (14,042 test and 1,531 validation).
- **OpenBookQA** (Mihaylov et al. 2018): Tests elementary science knowledge with 4,957 training, 500 validation, and 500 test samples.
- **BoolQ** (Clark et al. 2019; Wang et al. 2019): Dichotomous QA (yes/no) dataset with 9,427 training, 3,270 validation, and 3,245 test samples.
- **SciQ** (Johannes Welbl 2017): Contains 13,679 science questions (11,679 training, 1,000 validation, 1,000 test); answer order is randomized to prevent ordinal biases.
- **ARC (Challenge & Easy)** (Clark et al. 2018): The Challenge set contains 2,590 hard questions; the Easy set includes 5,197 questions from grade-school science exams.
- **MathQA** (Amini et al. 2019): Tests mathematical problem-solving using 4,475 validation samples.

4.3 Abstractive QA

For Abstractive QA, we evaluate the ability of LLMs to generate free-text answers without explicit candidates:

- **SQuADv2** (Rajpurkar et al. 2016; Rajpurkar, Jia, and Liang 2018): Repurposed for abstractive QA by conditioning the model solely on the transformed question.
- **TruthfulQA** (Lin, Hilton, and Evans 2022): The model generates free-text answers scored for semantic similarity with correct options.
- **WikiQA** (Yang, Yih, and Meek 2015): Adapted for abstractive tasks, with correctness determined by a cosine similarity score greater than 60% between the generated answer and labeled passages.
- **SciQ** (Johannes Welbl 2017): Evaluated without candidate choices, using approximate Levenshtein distance to match generated answers.
- **HotpotQA** (Yang et al. 2018): A Wikipedia-based QA dataset with 57,711 train and 5,600 validation samples, emphasizing factual, diverse topics.
- **TriviaQA** (Joshi et al. 2017): Contains 95,000 QA pairs authored by trivia experts, evaluated on 67,469 training and 11,313 validation samples.

5 Metrics

5.1 Accuracy, Robustness, & Plurality (S)

To comprehensively evaluate the performance of `gpt-3.5-turbo`, we analyze several key aspects: accuracy, robustness, and plurality-based voting under adversarial question perturbations. Since the original question q_0 is always included in our answer set $\mathcal{A}$, we juxtapose the baseline accuracy with ensemble metrics that capture worst-case performance, best-case outcomes, and agreement across perturbed answers. Inspired by prior LLM evaluations (Liang et al. 2022), our measure of robustness considers the following:

- **Worst-case Robustness** (**Ω**): Measures the lower bound, where at least one perturbed input is incorrect.
- **Best-case Robustness** (**O**): Measures the upper bound, where at least one perturbed input is correct.

For consistency, let m denote an indicator function for accuracy, A represent the baseline accuracy over n samples, and $\mathcal{T}_j(x_j)$ describe perturbations applied to input x_j. Let $\hat{Y}$ represent the ensemble answer generated by **plurality voting** over $v+1$ raters. The metrics are defined as follows:

$$m(f(x), y) = \mathbb{1}_{f(x)=y}$$

$$A = \frac{1}{n}\sum_{j=1}^{n} m(f(x_j), y_j)$$

$$\Omega = \frac{1}{n}\sum_{j=1}^{n} \min_i m\left(f(T_i(x_j)), y_j\right) \leq A$$

$$O = \frac{1}{n}\sum_{j=1}^{n} \max_i m(f(T_i(x_j)), y_j) \geq A$$

$$\hat{Y} = \frac{1}{n}\sum_{j=1}^{n} m\left(\text{mode}\{\, f(T_i(x_j)) \;\middle|\; T_i \in \mathcal{T}\,\}, y_j\right)$$

Where:

- $T_0(x) = I$: the identity function representing the original unperturbed question.
- $f(x)$: the model's output for input x.
- y: the ground truth answer.

Relationships Between Metrics The relationship between accuracy A, robustness (Ω, O), and plurality voting $\hat{Y}$ depends on the interaction between the original query and its perturbed variations:

$$\Omega \leq \min(\hat{Y}, A) \leq \max(\hat{Y}, A) \leq O$$

This hierarchy reflects that worst-case robustness (Ω) sets the lower bound, while best-case robustness (O) defines the upper bound. Perturbations can influence the model's performance and agreement in the following ways:

- *Perturbations Help Align Outputs*: Then $\hat{Y} \geq A$ as the mode benefits from consensus across the outputs.
- *Perturbations Introduce Noise*: Then $\hat{Y} \leq A$ as the mode is skewed by incorrect answers from variations.
- *Perturbations Are Neutral*: In this case, the model performs consistently across all queries ($\hat{Y} \approx A$).

Random Guessing: If the raters randomly guess among k possible answer choices, the following behaviors are observed:

- Accuracy A and plurality $\hat{Y}$ are $1/k$.
- Worst-case robustness (Ω) asymptotically approaches:

$$\lim_{k\to\infty} \left(\frac{1}{k}\right)^{v+1} \approx 0$$

- Best-case robustness (O) approaches:

$$1 - \left(\frac{k-1}{k}\right)^{(v+1)}$$

The best-case probability represents the likelihood of at least one correct answer across $v+1$ independent guesses.

Insights These metrics allow MultiQ&A to quantify the stability and robustness of LLMs when subjected to extensive adversarial scenarios:

- **Worst-case Robustness** captures the model's vulnerability to adversarial perturbations.
- **Best-case Robustness** highlights the model's potential for correctness under diverse perturbations.
- **Plurality Voting** provides a practical aggregate for robust decision-making, reflecting the consensus across multiple perturbed answers.

5.2 Agreement

Item Difficulty (S) Item difficulty μ_D measures how challenging each question is for the LLM raters by computing the average correctness across all responses (Lord 1952):

$$\mu_D = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{|\mathcal{T}|}\sum_{T_i\in\mathcal{T}} m(f(T_i(x_j)), y_j)$$

where n represents the number of samples, $\mathcal{T}$ is the set of perturbations, and m is the indicator function for correctness. For random guessing, the expected value follows a Bernoulli distribution, $\mathbb{E}\left[\mu_D\right] = 1/k$.

Mean Normalized Certainty (U) Entropy H quantifies uncertainty in the responses: higher entropy corresponds to greater uncertainty, while lower entropy reflects more consistent answers (Shannon 1948; Wilcox 1973). We normalize the rater entropy H by the maximum possible entropy H_{max}, inverting the scale to reflect certainty: 1 represents high certainty and 0 indicates uncertainty:

$$\begin{aligned} p_i &= \frac{f_i}{v+1} \\ H_\eta &= 1 - \mathbb{E}\left[\frac{H}{H_{max}}\right] \\ &= 1 + \frac{1}{n}\sum_{j=1}^{n}\left[\sum_{i=0}^{K_j}\frac{p_i \log_b(p_i)}{\log_b(K_j)}\right] \end{aligned}$$

where $H_\eta \in [0,1]$, f_i is the frequency of answer choice i, p_i is the proportion of answer choice i across $v+1$ raters, and K_j is the number of possible choices for question q_j.

Gibbs' M2 Index (U) The M_2 index quantifies the variance in rater's responses, assuming a multinomial distribution for the answer choices (Gibbs and Poston 1975). The index is standardized such that $M_2 = 1$ indicates complete certainty (no variability), while $M_2 = 0$ indicates a uniform distribution (maximum uncertainty):

$$M_2 = 1 - \frac{1}{n}\sum_{j=1}^{n}\left[\frac{K_j}{K_j-1}\left(1 - \sum_{i=0}^{K_j} p_i^2\right)\right]$$

Fleiss's Generalized κ (U) Fleiss' κ measures the degree of inter-rater agreement beyond what would be expected by random chance. A value of 1 indicates perfect agreement, while 0 indicates no agreement beyond chance (Cohen 1960; Fleiss 1971). Let f_i represent the frequency of answer choice k_i for sample x_j. The expected agreement by chance $\bar{P}_e$ and observed agreement $\bar{P}_o$ for $v+1$ raters is as follows:

$$\begin{aligned} \bar{P}_e &= \sum_{i=0}^{K_j}\left(\frac{1}{n(v+1)}\sum_{j=1}^{n} f_i\right)^2 \\ P_j &= \frac{1}{v(v+1)}\left[\left(\sum_{i=0}^{K_j} f_i^2\right) - (v+1)\right] \\ \bar{P}_o &= \frac{1}{n}\sum_{j=1}^{n} P_j \\ \kappa &= \frac{\bar{P}_o - \bar{P}_e}{1 - \bar{P}_e} \end{aligned}$$

where f_i is the frequency of answer choice k_i for each sample x_j, and K_j is the number of categories. Note that κ is affected by the number of raters and answer categories, with fewer categories often yielding higher κ values.

5.3 Reliability (S)

Cronbach's α measures the internal consistency and reliability of dichotomous responses (correct/incorrect) (Cronbach 1951). It is widely accepted in testing theory and is equivalent to the Kuder-Richardson Formula 20 (KR-20) for binary data (Kuder and Richardson 1937). The formula is:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n}\sigma_y^2}{\sigma_x^2}\right)$$

where n is the number of samples, σ_y^2 is is the variance in scores across $v+1$ raters for each sample, and σ_x^2 is the variance in total correct responses per rater.

6 Analysis & Discussion

As shown in Table 4, the performance of the different question answering formats across perturbations is: **Extractive** > **Multiple Choice** > **Abstractive** (Table 5). This ranking suggests that **Extractive** tasks outperform the others, due to the inherent advantage of additional content, such as context or answer choices, improving model robustness in Retrieval-Augmented scenarios (Lewis et al. 2021). **Multiple Choice** tasks benefit from the fixed set of choices, which provide some constraints that help guide the model's decision-making process. In contrast, **Abstractive** tasks experience greater variability in rater responses under perturbations, possibly due to the model's increased likelihood of hallucination or misinterpretation when generating free-form answers.

6.1 Extractive QA

Performance When provided with context, MultiQ&A demonstrates that `gpt-3.5-turbo` performs strongly on SQuADv2. In particular, the baseline and mode accuracy for the model are nearly identical, indicating that the model is quite stable under perturbations. Despite this high accuracy, adversarial question generation still manages to cause a failure rate of **13%** in the training set and **9.4%** in the validation

Datasets				Robustness				Agreement				Rel
	Name	**Split**	**#**	**Base**	**Mode**	**Worst**	**Best**	$\mu_\mathbf{D}$	$\mathbf{H}_\eta$	$\mathbf{M_2}$	κ	α
Extn	SQuADv2	train	80,049	91.9	90.8	68.6	97.3	87.0	85.8	84.4	75.0	99.9
		val	5,843	95.2	94.1	74.5	98.8	90.6	81.2	82.7	79.3	98.3
Multiple Choice	TruthfulQA	val	786	60.4	60.3	39.8	76.1	58.5	88.1	79.5	72.6	37.8
	PIQA	train	15,677	81.2	82.3	56.7	94.0	78.8	79.1	77.6	65.1	96.8
		val	1,784	80.1	83.2	58.2	93.8	79.2	79.3	77.9	66.0	82.7
	MMLU	dev	281	66.2	61.9	35.6	82.6	58.9	74.8	68.7	60.6	74.4
		val	1,463	64.5	65.0	37.9	82.8	60.3	74.8	68.8	60.8	85.9
		test	13,545	67.6	67.3	38.4	84.2	61.6	75.0	69.0	61.1	99.1
	OpenBook QA	train	4,909	78.0	75.6	37.8	91.3	67.8	72.8	66.6	58.0	99.1
		val	497	78.1	78.9	39.8	93.0	70.5	73.4	67.5	59.2	87.1
		test	499	75.6	73.7	38.1	90.8	66.8	73.1	66.9	58.4	88.9
	BoolQ	train	9,401	71.0	71.2	32.7	92.6	67.0	51.2	54.3	43.1	97.0
		val	3,256	71.5	71.7	33.5	93.2	67.6	51.5	54.7	43.3	91.6
	SciQ	train	11,670	93.4	92.9	76.7	97.6	89.9	91.7	89.4	86.4	98.7
		val	999	91.6	93.0	77.3	97.3	89.9	91.7	89.6	86.7	45.5
		test	998	93.8	93.5	76.9	97.8	90.6	91.9	89.8	87.0	85.6
	ARC - Challenge	train	1,118	85.5	82.9	52.3	95.2	76.8	82.7	76.9	69.9	95.6
		val	299	87.0	82.3	53.8	95.0	77.1	82.5	76.4	68.9	88.9
		test	1,172	83.8	80.7	50.9	92.7	74.7	81.1	76.3	70.3	96.1
	ARC - Easy	train	2,248	93.3	92.7	70.8	97.9	88.0	90.1	87.4	83.7	96.6
		val	570	94.4	92.1	66.1	98.1	86.7	87.6	84.6	80.8	92.4
		test	2,374	92.8	92.5	69.6	98.3	87.9	90.2	86.8	82.9	95.7
	MathQA	train*	693	50.1	56.9	9.5	85.9	46.6	60.1	46.1	30.2	39.0
		val	4,473	49.8	55.5	9.4	85.8	45.9	64.4	47.8	29.8	89.7
		test	2,985	47.7	54.7	9.2	84.5	45.6	68.1	50.0	31.3	45.2
Abstractive	SQuADv2	train	27,206	32.9	31.8	15.1	46.7	29.9	74.7	76.4	66.3	98.5
		val	5,864	25.4	24.0	10.2	37.9	22.9	90.4	86.3	64.6	94.3
	TruthfulQA	val	807	52.4	28.1	61.8	78.6	55.1	58.9	61.5	53.4	31.3
	WikiQA	train	1,028	73.4	72.6	54.6	80.9	69.8	79.6	81.3	77.6	86.5
		val	140	76.4	75.7	58.6	82.9	73.3	80.9	82.4	78.3	35.6
		test	286	73.8	69.9	52.4	81.1	67.8	76.8	78.4	74.0	80.6
	SciQ	train	11,596	66.2	75.6	35.3	80.4	59.8	80.6	74.7	69.0	99.2
		val	991	65.7	74.7	34.8	80.1	58.9	80.6	74.8	68.9	91.0
		test	995	70.1	77.4	36.8	83.2	62.4	80.5	74.6	69.1	93.5
	HotpotQA (KILT)	train	66,345	45.5	41.7	21.2	59.6	40.6	80.7	78.6	65.8	99.8
		val	5,542	42.5	38.8	21.5	55.9	38.3	72.5	74.4	69.3	96.8
	TriviaQA	train	76,635	71.7	69.5	48.6	79.7	66.8	84.6	83.1	72.8	99.9
		dev	11,177	72.2	69.8	48.5	80.1	67.0	84.8	82.8	72.5	99.1

Table 4: Experimental results for **Extractive**, **Multiple Choice**, and **Abstractive** QA scenarios across each dataset split and metrics for accuracy (Base), plurality (Mode), robustness (Worst & Best), item difficulty (μ_D), agreement (Fleiss's κ, H_η, M_2), and reliability (Cronbach α). For MathQA train, we evaluated 693 out of 29,800 samples.

Scenario			Robustness				Agreement				Rel
	Experiment	**#**	**Base**	**Mode**	**Worst**	**Best**	$\mu_\mathbf{D}$	$\mathbf{H}_\eta$	$\mathbf{M_2}$	κ	α
AVG	**Extractive**	85,892	93.6	92.5	71.6	98.1	88.8	83.5	83.6	77.2	99.1
	Multiple Choice	81,697	76.4	76.6	46.6	91.3	71.2	77.2	71.9	63.3	83.0
	Abstractive	208,612	59.1	57.7	38.4	71.3	54.8	78.9	77.6	69.4	79.6
	Total	376,201	71.4	70.9	45.1	84.8	66.5	78.1	74.4	66.1	82.7
	Experiment	**#**	**Base**	**Mode**	**Worst**	**Best**	$\mu_\mathbf{D}$	$\mathbf{H}_\eta$	$\mathbf{M_2}$	κ	α
WAVG	**Extractive**	85,892	92.1	91.0	69.0	97.4	87.2	85.5	84.3	75.3	99.8
	Multiple Choice	81,697	76.3	76.8	47.4	91.6	71.8	75.2	71.3	61.9	92.5
	Abstractive	208,612	55.9	53.9	32.9	67.3	51.2	81.5	80.0	69.1	98.8
	Total	376,201	68.6	67.4	44.3	79.4	63.9	81.0	79.1	69.0	97.7

Table 5: Aggregated results by scenario and in total, displaying unweighted averages (AVG) and weighted averages (WAVG) to reduce bias from larger datasets and enable a holistic analysis across all knowledge domains.

set. This suggests that, although the model performs well on a majority of questions, there is a subset where it struggles, especially when the question formulations are intentionally altered.

Agreement Moreover, `gpt-3.5-turbo` exhibits **75%** agreement among perturbations and **79.3%** agreement on the validation set. This level of agreement can be interpreted as substantial, indicating that the perturbations applied did not significantly compromise the model's ability to provide consistent answers (Landis and Koch 1977). The LLM's performance remains resilient under perturbations, especially when sufficient context is provided, enabling it to correctly answer questions even with with significant prompt alterations.

Worst-Case Performance The worst-case performance of `gpt-3.5-turbo` is particularly noteworthy, as it outperforms all other experiments. This robustness highlights that when context is available, the model can generally handle a wide range of perturbations and still produce correct answers, reaffirming the importance of context in mitigating potential performance degradation due to adversarial inputs.

6.2 Multiple Choice QA

Robustness Most variations in the **Multiple Choice** format benefit from the presence of fixed answer choices to rely on, demonstrating significant robustness across five perturbations. The model's reliance on unaltered answer choices helps maintain consistency, even when the question is modified. We see that `gpt-3.5-turbo` exhibits strong performance across several benchmarks, achieving **67.6%** accuracy on MMLU in a 0-shot setting, **60.4%** accuracy on TruthfulQA, and **83.8%** (C) and **92.8%** (E) on the ARC Challenge and Easy sets, respectively. These results highlight the model's capability to handle a variety of multiple-choice questions with a solid level of accuracy. For other benchmarks, the model shows varied performance: **80.1%** on PIQA, **71.5%** on BoolQ, and **72.2%** on TriviaQA in a 0-shot environment. While the model performs well on most multiple-choice tasks, the differences in its performance across datasets emphasize the importance of dataset characteristics and question types that play a significant role in influencing accuracy.

Ensemble Accuracy & Internal Consistency Our experiments focused on evaluating robustness under perturbations rather than conducting few-shot ablation studies. For most datasets, `gpt-3.5-turbo` achieves high ensemble accuracy, consistent with the baseline, suggesting that perturbations do not significantly disrupt the most frequent response. Internal consistency across raters is also very high, likely due to the strong agreement and accuracy, indicating that the model remains stable even under slight changes in the question formulation.

Outlier: MathQA However, **MathQA** stands out as an outlier in our analysis. The model demonstrates low accuracy, poor worst-case performance, and a low μ_D, suggesting that the questions in MathQA are particularly challenging for `gpt-3.5-turbo`. Furthermore, the agreement between raters is low at approximately **30%**, indicating that the model struggles to provide consistent answers across different perturbations. This is corroborated by the low consistency scores (Cronbach's $\alpha \approx 42.8\%$ on the test), signaling to `gpt-3.5-turbo`'s difficulties in performing arithmetic operations within a language modeling framework (Mirzadeh et al. 2024).

6.3 Abstractive QA

Performance Without Context In the **Abstractive** format, `gpt-3.5-turbo` performs well even without the inclusion of additional context, demonstrating notable robustness at a temperature setting of $\tau = 1.0$ and under $v = 5$ perturbations. The model's ability to handle perturbations in the absence of explicit context suggests a flexible approach to generating answers. However, the accuracy still fluctuates depending on the difficulty of the question and the ground truth, indicating that while the model is adaptive, its performance is sensitive to task complexity.

Variation Across Tasksets Significant variation is observed in accuracy across different tasksets. Specifically, tasksets like TriviaQA, SciQ, and WikiQA show an improvement of **+20%** in accuracy compared to SQuADv2, TruthfulQA, and HotpotQA. This highlights that simpler tasksets, with more straightforward ground truths, are easier for generative models like `gpt-3.5-turbo` to answer. The results suggest that the complexity of the task and the nature of the questions play a critical role in the model's performance, with simpler or more direct questions yielding better outcomes in **Abstractive** settings.

6.4 Abusive or Sensitive Content

We encountered **2,293** cases where `gpt-3.5-turbo` failed to generate responses due to content filtering, separate from standard service errors (`APIError`, `ServiceUnavailableError`, `RateLimitError`). These failures fell into two main categories:

- `AttributeError` (**1,081** cases): Triggered when generating violent or explicit content.
- `InvalidRequestError` (**1,212** cases): Occurs when prompt filtering flags violent or explicit terms in the input.

These cases can enrich adversarial datasets for content filtering. Below is the observed distribution across datasets:

MMLU	638	PIQA	543	SQuADv2	490
TriviaQA	326	HotpotQA	103	BoolQ	57
OpenBookQA	48	TruthfulQA	29	SciQ	25
WikiQA	19	MathQA	15		

6.5 Token Usage and Statistics

Our evaluation spanned 376,201 questions, producing 1,881,005 variations and 2,257,206 total answers. The process utilized 717,530,842 tokens, including 115,834,262 for perturbations and 601,696,580 for answer generation.

7 Conclusion

MultiQ&A is a crowdsourcing-based method to assess the robustness and consistency of LLM-generated answers. By perturbing 376,201 questions into 1,881,005 lexical variations while preserving semantics, our experiments across 13 datasets quantified consistency, reliability, and robustness, providing valuable insights into LLM responses. We believe that MultiQ&A provides a promising infrastructure for institutions adopting LLMs with increased confidence.

References

Amini, A.; Gabriel, S.; Lin, S.; Koncel-Kedziorski, R.; Choi, Y.; and Hajishirzi, H. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2357–2367. Minneapolis, Minnesota: Association for Computational Linguistics.

Azaria, A.; and Mitchell, T. 2023. The Internal State of an LLM Knows When It's Lying. arXiv:2304.13734.

Bisk, Y.; Zellers, R.; Bras, R. L.; Gao, J.; and Choi, Y. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Cho, N.; Srishankar, N.; Cecchi, L.; and Watson, W. 2024. FISHNET: Financial Intelligence from Sub-querying, Harmonizing, Neural-Conditioning, Expert Swarms, and Task Planning. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, 591–599. New York, NY, USA: Association for Computing Machinery. ISBN 9798400710810.

Clark, C.; Lee, K.; Chang, M.-W.; Kwiatkowski, T.; Collins, M.; and Toutanova, K. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *NAACL*.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457v1*.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3): 297–334.

Dohan, D.; Xu, W.; Lewkowycz, A.; Austin, J.; Bieber, D.; Lopes, R. G.; Wu, Y.; Michalewski, H.; Saurous, R. A.; Sohl-dickstein, J.; Murphy, K.; and Sutton, C. 2022. Language Model Cascades. arXiv:2207.10342.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382.

Gibbs, J. P.; and Poston, J., Dudley L. 1975. The Division of Labor: Conceptualization and Related Measures*. *Social Forces*, 53(3): 468–476.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Izacard, G.; and Grave, E. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 874–880. Online: Association for Computational Linguistics.

Johannes Welbl, M. G., Nelson F. Liu. 2017. Crowdsourcing Multiple Choice Science Questions.

Johnson, J.; Douze, M.; and Jégou, H. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3): 535–547.

Joshi, M.; Choi, E.; Weld, D.; and Zettlemoyer, L. 2017. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, arXiv:1705.03551.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. Online: Association for Computational Linguistics.

Khashabi, D.; Min, S.; Khot, T.; Sabharwal, A.; Tafjord, O.; Clark, P.; and Hajishirzi, H. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1896–1907. Online: Association for Computational Linguistics.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916.

Kuder, G.; and Richardson, M. 1937. The theory of the estimation of test reliability. *Psychometrika*, 2(3): 151–160.

Kumar, S.; Paria, B.; and Tsvetkov, Y. 2022. Gradient-based Constrained Sampling from Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2251–2277. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Landis, J. R.; and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159–174.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

Li, Y.; Lin, Z.; Zhang, S.; Fu, Q.; Chen, B.; Lou, J.-G.; and Chen, W. 2023. Making Large Language Models Better Reasoners with Step-Aware Verifier. arXiv:2206.02336.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Ré, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2022. Holistic Evaluation of Language Models. arXiv:2211.09110.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252. Dublin, Ireland: Association for Computational Linguistics.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2021. What Makes Good In-Context Examples for GPT-3? arXiv:2101.06804.

Lord, F. M. 1952. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 17(2): 181–194.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9802–9822. Toronto, Canada: Association for Computational Linguistics.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Conference on Empirical Methods in Natural Language Processing*.

Mirzadeh, I.; Alizadeh, K.; Shahrokhi, H.; Tuzel, O.; Bengio, S.; and Farajtabar, M. 2024. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. arXiv:2410.05229.

Moradi, M.; and Samwald, M. 2021. Evaluating the Robustness of Neural Language Models to Input Perturbations. arXiv:2108.12237.

Nogueira, R.; and Cho, K. 2020. Passage Re-ranking with BERT. arXiv:1901.04085.

Nye, M.; Andreassen, A. J.; Gur-Ari, G.; Michalewski, H.; Austin, J.; Bieber, D.; Dohan, D.; Lewkowycz, A.; Bosma, M.; Luan, D.; Sutton, C.; and Odena, A. 2021. Show Your Work: Scratchpads for Intermediate Computation with Language Models. arXiv:2112.00114.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. arXiv:1806.03822.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv:1606.05250.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423.

Slobodkin, A.; Goldman, O.; Caciularu, A.; Dagan, I.; and Ravfogel, S. 2023. The Curious Case of Hallucinatory (Un)answerability: Finding Truths in the Hidden States of Over-Confident Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3607–3625. Singapore: Association for Computational Linguistics.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv preprint 1905.00537*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171.

Watson, W.; Cho, N.; Balch, T.; and Veloso, M. 2023. HiddenTables and PyQTax: A Cooperative Game and Dataset For TableQA to Ensure Scale and Data Privacy Across a Myriad of Taxonomies. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7144–7159. Singapore: Association for Computational Linguistics.

Watson, W.; Cho, N.; Srishankar, N.; Zeng, Z.; Cecchi, L.; Scott, D.; Siddagangappa, S.; Kaur, R.; Balch, T.; and Veloso, M. 2025. LAW: Legal Agentic Workflows for Custody and Fund Services Contracts. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, 583–594. Abu Dhabi, UAE: Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Wilcox, A. R. 1973. Indices of Qualitative Variation and Political Measurement. *The Western Political Quarterly*, 26(2): 325–343.

Wu, T.; Jiang, E.; Donsbach, A.; Gray, J.; Molina, A.; Terry, M.; and Cai, C. J. 2022. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391566.

Yang, Y.; Yih, W.-t.; and Meek, C. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2013–2018. Lisbon, Portugal: Association for Computational Linguistics.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380. Brussels, Belgium: Association for Computational Linguistics.

Yu, W.; Iter, D.; Wang, S.; Xu, Y.; Ju, M.; Sanyal, S.; Zhu, C.; Zeng, M.; and Jiang, M. 2023. Generate rather than retrieve: Large language models are strong context generators. In *International Conference for Learning Representation (ICLR)*.

Zeng, Z.; Watson, W.; Cho, N.; Rahimi, S.; Reynolds, S.; Balch, T.; and Veloso, M. 2023. FlowMind: Automatic Workflow Generation with LLMs. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, 73–81. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702402.

Zhang, Y.; Pan, L.; Tan, S.; and Kan, M.-Y. 2022. Interpreting the Robustness of Neural NLP Models to Textual Perturbations. arXiv:2110.07159.

A Disclaimer

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates (”JPMorgan”) and is not a product of the Research Department of JPMorgan. JPMorgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.