Indo American Journal of Multidisciplinary Research and Review (IAJMRR)
International Peer Reviewed - Refereed Research Journal
ISSN: 2581 - 6292, Impact Factor: 6.885, Website: www.iajmrr.com

ASSESSING ASSOCIATIONS BETWEEN BENCHMARK ITEM CHARACTERISTICS AND REFERENCE ANSWER CORRECTNESS IN LANGUAGE MODEL TRUTHFULNESS EVALUATION

Divya Vishwanath* & Paul Johnson Asamoah**
* Janardhan Rai Nagar Rajasthan Vidyapeeth, Udaipur, Rajasthan, India
** Power Life University College, Accra, Ghana, West Africa

Cite This Article: Divya Vishwanath & Paul Johnson Asamoah, "Assessing Associations Between Benchmark Item Characteristics and Reference Answer Correctness in Language Model Truthfulness Evaluation", Indo American Journal of Multidisciplinary Research and Review, Volume 8, Issue 2, July - December, Page Number 166-177, 2024.
Copy Right: © IAJMRR Publication, 2024 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract:

This analysis examines how reference composition shapes the interpretation of truthfulness benchmarks used in international language model evaluation. A descriptive cross-sectional analysis uses 6,202 reference answers nested within 817 questions across 38 categories from the 2021. Counts, correctness proportions and answer length distributions are compared using pooled weighting, equal question weighting and contrasts within questions. Adversarial questions contain a higher correct reference proportion, but the difference decreases from 4.80 to 3.35 percentage points when questions receive equal weight. Category proportions range from 21.59 to 60.71 percent. Correct references average 0.99 words longer than incorrect references, while the average contrast within questions is 0.73 words. Opposing patterns across questions and categories qualify this aggregate tendency. Unequal reference enumeration changes descriptive magnitudes by changing each question's contribution to pooled summaries. The analysis refines benchmark validity by distinguishing reference composition from evaluated capability. Benchmark maintainers and evaluation teams should disclose question denominators, category coverage and weighting conventions to support transparent interpretation across institutional settings.

Key Words: Answer Length; Benchmark Validity; Descriptive Analysis; Reference Composition; Truthfulness Evaluation

1. Introduction:

Truthfulness evaluation depends on how benchmarks represent both correct answers and plausible falsehoods. A reference corpus supplies the alternatives against which responses can be interpreted, yet its composition requires scrutiny in its own right. Questions may contribute different numbers of references, substantive categories may receive unequal coverage, and correct and incorrect answers may differ in textual form. These features affect what aggregate summaries describe. For researchers and institutions using shared benchmarks, the substantive problem is therefore how to distinguish the composition of an evaluation resource from the capability that resource is intended to assess. TruthfulQA provides a relevant setting for examining this distinction. Lin et al., 2022, introduced the benchmark to evaluate whether language models reproduce falsehoods associated with common misconceptions. Its questions are accompanied by correct and incorrect reference answers, question classifications and topic categories. The benchmark therefore supports an analysis of how labelled references are distributed across observable item characteristics. Reference answer correctness has a specific meaning in this setting: it records the classification supplied by the benchmark authors. A proportion calculated from these labels describes the reference pool, whereas a model accuracy measure would require observed model responses.

The theoretical basis for examining reference composition lies in the distinction between an intended evaluation construct and the statistical regularities through which it is represented. Geirhos et al., 2020, explain how learning systems can exploit shortcuts that perform well under particular evaluation conditions without capturing the intended relationship. At the dataset level, Gururangan et al., 2018, demonstrate that annotation processes can introduce features associated with labels in natural language inference. Together, these contributions establish that observable regularities deserve examination before they receive a substantive interpretation. They do not establish that a particular regularity exists in TruthfulQA or that a model exploits it. Three observed characteristics provide a focused basis for that examination. Question type distinguishes adversarial from nonadversarial questions and may accompany differences in the composition of reference alternatives. Topic category identifies the substantive domain and allows comparisons across questions with different factual content. Answer length captures textual extent, which may differ between concise assertions, qualified explanations and elaborated misconceptions. These characteristics need not have uniform relationships with correctness. Correct references may require qualifications, but incorrect references may also be lengthy. Bender and Koller, 2020, distinguish linguistic form from meaning, reinforcing the need to interpret word count as a textual measure rather than evidence of understanding or factual adequacy.

Aggregation creates a further measurement issue because reference answers are nested within questions. A pooled correctness proportion gives each answer equal weight and consequently gives greater influence to questions containing more references. Averaging correctness proportions across questions instead gives each question equal influence. Both calculations are legitimate, but they answer different descriptive questions. Paullada et al., 2021, argue that dataset analysis should examine development practices and the conditions of subsequent use. Applied here, that perspective makes the weighting convention part of the interpretation rather than a reporting detail. The closest literature establishes the purpose of truthfulness benchmarking, the possibility of annotation artifacts and the importance of dataset scrutiny. The reviewed evidence does not establish how question type, topic category and answer length jointly characterise the selected TruthfulQA reference pool, or how unequal answer counts change its descriptive summaries. This analysis addresses that bounded question through a reproducible examination of one benchmark version. Its contribution concerns the separation of reference composition, textual variation and aggregation. It does not depend on asserting that these features identify a new causal mechanism.

This study aims to characterise reference answer correctness across question types, topic categories and answer lengths within the selected TruthfulQA snapshot. The central research question asks how these characteristics relate descriptively to correctness labels and how the resulting patterns change when questions receive equal weight. The objectives are to compare correctness composition between question types, document variation across topic categories, and examine answer length distributions across correctness classes. Comparisons within questions additionally assess whether pooled length differences persist when the question provides the common context. The empirical analysis uses a cross section of 6,202 retained reference answers nested within 817 questions across 38 categories from the 2021 repository snapshot. It preserves source labels, excludes conflicting label combinations and retains question identifiers throughout the comparisons. Counts, proportions and length distributions address the research objectives directly. Comparing pooled summaries with averages calculated across questions makes the influence of unequal reference counts visible without requiring assumptions about an external sampling population.

The findings identify a higher correct reference proportion among adversarial questions, substantial variation across categories and longer correct references on average. However, weighting changes the magnitudes, and the length relationship reverses in some questions and categories. These results qualify an interpretation based solely on pooled averages. The contribution to benchmark validity is the demonstration that reference composition, the unit receiving analytical weight and textual heterogeneity must be considered together. For benchmark maintainers and evaluation researchers, the practical implication is to preserve question identifiers and report category denominators and weighting conventions alongside aggregate summaries. This reporting principle may inform other reference-based evaluations, while the observed magnitudes remain specific to the English language benchmark version examined.

2. Theory, Conceptual Framework and Research Questions Development:

This analysis draws on theories of benchmark validity and shortcut learning to examine how reference answer correctness varies with observable item characteristics. Benchmark developers select questions, construct answer alternatives and assign correctness labels. These decisions jointly shape the relationships available for analysis. Lin et al., 2022, introduced TruthfulQA to assess responses to questions that can elicit common misconceptions. The available dataset contains its reference answers and author supplied labels. Accordingly, the analytical focus is the composition of the reference answer pool. Benchmark validity requires distinguishing the intended construct from observable features that accompany it. Geirhos et al., 2020, explain that predictive systems can exploit convenient statistical regularities rather than the relationships intended by an evaluation task. This perspective motivates examining question type, topic category and answer length as possible sources of variation in reference labels. These characteristics are observed variables, not indicators of a latent scale. Their relationships with correctness remain empirical questions rather than established findings.

Question type distinguishes adversarial from nonadversarial questions. Adversarial selection targets prompts associated with plausible misconceptions, which may influence the kinds and numbers of incorrect alternatives included during benchmark construction. However, this selection mechanism does not determine the proportion of correct reference answers. Developers can provide several valid formulations alongside several incorrect alternatives. The expected direction therefore depends on how the reference answer pool was assembled. Evidence from other language benchmarks demonstrates why construction processes deserve attention. Gururangan et al., 2018, identified annotation artifacts that allowed inference labels to be predicted from features introduced during data creation. Their findings establish that benchmark construction can generate label associations, but do not establish the direction of an association in TruthfulQA. This analysis therefore describes the numbers and proportions of correct and incorrect references within each question type.

RQ1: How do the numbers and proportions of correct reference answers differ between adversarial and nonadversarial questions?

Topic category captures the substantive domain of each question. Unlike question type, which records a selection distinction, category groups questions by their content. Domains may differ in the specificity of acceptable answers, the availability of alternative formulations and the misconceptions represented in incorrect references. These differences provide a conceptual basis for examining category variation without treating categories as an ordered measure. Category differences can also reflect unequal question coverage and answer enumeration. A category containing more reference answers per question contributes more heavily to an answer level proportion. Niven and Kao, 2019, demonstrated that apparent performance on an argument comprehension benchmark could depend on superficial dataset cues. Their evidence supports examining benchmark composition before attributing observed differences to substantive capability. Here, category comparisons describe reference labels and retain the question as the grouping unit.

RQ2: How does the proportion of correct reference answers vary across topic categories, alongside differences in question coverage and reference answer counts?

Answer length records the number of words in each reference answer. Longer answers may express qualifications, specify conditions or correct a misconception, making them compatible with accurate reference formulations. Conversely, incorrect alternatives may also contain elaboration, while correct answers may require only a short factual statement. These competing mechanisms provide no defensible directional prediction. Bender and Koller, 2020, distinguish linguistic form from meaning and argue against treating formal regularities as sufficient evidence of understanding. This distinction limits what answer length can represent. Word count measures textual extent rather than factual adequacy or reasoning quality. The analysis therefore compares length distributions across correctness labels and examines whether their pattern differs across question types and topic categories.

RQ3: How do reference answer length distributions differ between correct and incorrect answers, overall and within question types and topic categories?

The conceptual framework links three observed characteristics, question type, topic category and answer length, to the observed correctness label through descriptive comparisons. Correctness equals one for references listed as correct and zero for references listed as incorrect. The analytical dataset contains 6,202 reference answers nested within 817 questions. Question identifiers preserve this dependence when summarising patterns. Question type and topic category describe the question, whereas answer length and correctness describe individual references. These relationships represent associations within a curated benchmark. The descriptive analysis reports counts, proportions and length distributions without imposing directional hypotheses. It presents both pooled answer proportions and the distribution of correctness proportions across questions, since these summaries give different weight to questions with unequal numbers of references. The framework allows similarity, divergence or an unclear pattern across groups without requiring a predetermined result.

3. Data:

This analysis documents the benchmark source, reconstructs the analytical sample and defines the observed variables. It uses descriptive comparisons to examine reference answer correctness across question types, topic categories and answer lengths.

3.1 Data Sources, Population, and Sample Construction:

The data originate from TruthfulQA, an English language benchmark developed by Stephanie Lin, Jacob Hilton and Owain Evans. The benchmark contains questions and reference answers designed to support truthfulness evaluation, as documented by Lin et al., 2022. The snapshot provides 817 questions, their classifications, correct and incorrect reference answers, and supporting source links. The repository distributes the material under the Apache License, version 2.0. The accompanying workbook records the retrieval date and preserves the source records. The analytical evidence comes from the 2021 snapshot within the eligibility window of 2010 to 2023. That window defines source eligibility rather than repeated annual observations.

The target population is the reference answer pool contained in the selected benchmark version. The accessible sampling frame comprises all nonempty entries in its Correct Answers and Incorrect Answers fields. The selection approach is exhaustive within this frame. It does not constitute probability sampling from all questions, factual statements or language model responses. The observation unit is a reference answer nested within a question. The resulting structure is a cross section with clustered observations. It covers 38 topic categories without a geographical sampling restriction or a sector specific sampling design. These categories support comparisons across substantive domains, but they do not represent national or sectoral populations. Sample construction retained all 817 questions. Splitting the answer fields produced 6,209 nonempty entries. Removing five repeated entries within questions left 6,204 unique question and answer combinations. Two combinations appeared under both correctness labels and were excluded, producing 6,202 retained observations.

Stage	Questions Retained	Answer Observations	Reason or Rule
Accessible question	817	Not yet expanded	All questions in the selected snapshot

Stage	Questions Retained	Answer Observations	Reason or Rule
frame			
Expanded answer frame	817	6,209	Nonempty entries from both answer fields
Duplicate removal	817	6,204	Five repeated entries collapsed within questions
Conflicting label exclusion	817	6,202	Two unique answers assigned both labels removed
Final descriptive sample	817	6,202	Complete coverage of the analytical variables

The final sample contains 437 adversarial questions with 3,420 reference answers and 380 nonadversarial questions with 2,782 reference answers. Reference counts range from two to 24 per question, with a median of seven. Consequently, pooled answer summaries assign greater weight to questions containing more reference alternatives.

3.2 Data Integration, Cleaning, and Quality Assurance:

Data preparation expanded one source rather than merging independent datasets. Each question received a sequential identifier linked to its position in the source file. Its question type, category and question text were carried into every associated answer record. This created one question record linked to multiple answer records while preserving the original fields in the workbook. Correct and incorrect answer fields were split at semicolons, and surrounding whitespace was removed. Exact answer strings were compared within each question. Identical strings associated with different questions remained separate because their meaning and correctness depend on the question context. The source labels supplied the outcome coding; the preparation process did not introduce new factual judgments.

Five repeated entries occurred across five questions. Three represented repetition without a label conflict. The other two concerned the answer Unknown in source question records 348 and 355, where the same string appeared in both correctness classes. Collapsing repeated strings and then excluding these two conflicting combinations reconciles the initial and retained observation counts. No question was removed. The analytical checks confirmed unique answer identifiers and agreement between recorded word counts and counts recomputed from answer text. All 6,202 retained records contain question identifiers, question types, categories, answer text, word counts and correctness labels. No imputation, interpolation, external matching or replacement of missing values with zero was applied.

Answer lengths range from one to 51 words. All observed lengths were retained without trimming or winsorisation. The preparation process treated length as a textual characteristic rather than defining extreme observations as errors. No currency adjustment, temporal aggregation or frequency conversion was required. The analysis uses a fixed historical snapshot rather than the current repository contents. This preserves the benchmark definitions and answer labels available in that version. Documentation of provenance, composition and preparation follows the transparency principles articulated by Gebru et al., 2021. The workbook preserves the source data, cleaned observations, variable dictionary and source identifiers. The cleaning record does not contain a dated preregistration or a comparison of alternative cleaning rules. The procedures are therefore reported as implemented preparation decisions. Although every question remains represented, the removal of conflicting answers defines the retained answer population and should remain visible when interpreting its composition.

3.3 Variable Measurement, Construction, and Validation:

Reference answer correctness is the binary outcome. An answer receives one when it originates from Correct Answers and zero when it originates from Incorrect Answers. The label records the benchmark authors' classification. It is not a score estimated from answer length, question type or category. Question type distinguishes adversarial and nonadversarial questions. Topic category preserves the original nominal classification across 38 domains. Answer length is the number of whitespaces separated tokens in the trimmed reference answer. These measures capture selection type, substantive domain and textual extent, respectively. They remain separate observed characteristics rather than components of a composite index.

Variable or Indicator	Analytical Role	Definition and Construction	Source and Identifier	Unit and Frequency	Transformation and Direction	Missing Value Rule	Coverage and Limitation
Reference answer correctness	Outcome	One for a correct reference and zero for an incorrect reference	TruthfulQA, Correct Answers and Incorrect Answers	Binary label per answer in the 2021 snapshot	Higher values identify the correct label	Conflicting labels excluded; none missing among retained records	6,202 answers; author supplied classification
Question type	Grouping variable	Original adversarial or nonadversarial classification	TruthfulQA, Type	Nominal category per question	No numerical ordering	None missing	Two groups; shared by answers within

Variable or Indicator	Analytical Role	Definition and Construction	Source and Identifier	Unit and Frequency	Transformation and Direction	Missing Value Rule	Coverage and Limitation
							each question
Topic category	Grouping variable	Original substantive category	TruthfulQA, Category	Nominal category per question	No numerical ordering	None missing	38 categories with unequal coverage
Answer length	Textual measure	Count of whitespace separated tokens after trimming surrounding whitespace	TruthfulQA reference answer text	Word count per answer	Higher values indicate longer text	None missing	6,202 answers; does not measure semantic complexity
Question identifier	Grouping identifier	Sequential identifier linked to the source record position	Source question record	Identifier per question	No substantive ordering	None missing	817 questions; preserves nesting
Answer identifier	Record identifier	Question identifier combined with the answer sequence	Constructed during expansion	Identifier per retained answer	No substantive ordering	None missing	6,202 unique identifiers
Release year	Provenance field	Benchmark snapshot year	Repository version	Constant value of 2021	No analytical transformation	None missing	Identifies the release rather than the date of each underlying fact

Validation addresses the consistency of observed measures. Word counts were checked against retained text, identifiers were checked for uniqueness, and sample totals were reconciled with the source expansion. These checks establish reproducible construction, rather than independent validation of every factual label. Bender and Koller, 2020, distinguish linguistic form from meaning, supporting the interpretation of word count as textual extent rather than understanding or factual adequacy. The sample contains 2,835 correct references and 3,367 incorrect references. The pooled correctness proportion is 45.71 percent. Answer length averages 8.70 words, with a standard deviation of 4.59 and a median of eight. Correct references average 9.23 words, while incorrect references average 8.24 words. These values describe the retained reference pool.

Measure	Valid Observations	Mean	SD	Median	Minimum	Maximum
Reference correctness	6,202	0.457	0.498	0	0	1
Answer length, all references	6,202	8.70	4.59	8	1	51
Answer length, correct references	2,835	9.23	4.91	9	1	51
Answer length, incorrect references	3,367	8.24	4.25	8	1	29
Retained answers per question	817	7.59	2.68	7	2	24

Category coverage ranges from four to 100 questions. At the answer level, category sizes range from 28 to 705 observations. This variation matters because annotation practices and the enumeration of reference alternatives can generate label patterns. Gururangan et al., 2018, demonstrated such annotation artifacts in natural language inference data. Their findings motivate separating the description of label composition from claims about substantive capability.

3.4 Analytical Framework, Identification, and Reproducibility:

The analytical framework is descriptive. It addresses how correctness labels and answer lengths are distributed across the observed benchmark characteristics. It does not estimate causal effects or require directional hypotheses. Counts and proportions address question type and category comparisons, while means, standard deviations, medians and ranges describe answer length. Equation 1 defines the pooled correctness proportion for a group.

$$p_g = \frac{\sum_{i \in A_g} y_i}{N_g}$$

Here, p_g is the correctness proportion in group g, A_g is the set of retained answers in that group, y_i is the correctness label of answer i, and N_g is the number of retained answers in the group. Multiplication by 100 expresses the proportion as a percentage. A group can represent the full sample, a question type or a topic category. Equation 2 gives each question equal weight.

$$\bar{p}_g = \frac{1}{Q_g} \sum_{q \in G_g} \frac{\sum_{i \in A_q} y_i}{n_q}$$

Here, $\bar{p}_g$ is the mean question specific correctness proportion in group g, Q_g is the number of questions in that group, G_g is its set of questions, $A_q$is the set of retained answers associated with question q, and $n_q$is the number of those answers. The outcome label $y_i$retains its definition from Equation 1. These summaries answer different descriptive questions. The pooled proportion describes the composition of all retained answers, whereas the mean question proportion describes the average composition across questions. For the complete sample, the respective values are 45.71 percent and 46.42 percent. Their

difference reflects unequal reference counts across questions. The comparison assesses the influence of weighting without changing the sample.

All comparisons preserve the nesting of answers within questions. The analysis does not treat 6,202 answers as independent draws from an external population or attach significance claims to descriptive differences. Category summaries retain their question-and-answer denominators so that differences in coverage remain visible. Interpretation is restricted to benchmark composition. Geirhos et al., 2020, explain how observable regularities can support shortcut learning without establishing the intended capability. Accordingly, associations involving answer length or question classifications describe potential regularities in the reference pool, rather than evidence that a language model uses them. Reproduction begins with the identified repository snapshot, expands the two answer fields, trims surrounding whitespace, collapses exact strings within questions, removes conflicting label combinations and recalculates word counts. The accompanying workbook preserves the original records and the resulting analytical variables. The same retained observations support both weighting approaches, ensuring that differences arise from aggregation rather than changes in eligibility.

4. Methods:

4.1 Research Design and Analytical Scope:

This analysis uses a descriptive cross sectional design to examine reference answer correctness across question types, topic categories and answer lengths. The observation unit is a retained reference answer nested within a benchmark question. The analytical sample contains 6,202 answers associated with 817 questions from the 2021 TruthfulQA snapshot. Table 1 documents sample construction, while Table 2 provides the measurement dictionary. The descriptive targets are the composition of the retained answer pool and the average composition of its constituent questions. These targets differ because questions contribute unequal numbers of references. The analysis therefore preserves question identifiers and distinguishes comparisons that weight answers equally from comparisons that weight questions equally.

The research questions determine the analytical sequence. Comparisons between adversarial and nonadversarial questions address differences in correctness composition. Category summaries examine variation across substantive domains alongside their coverage. Comparisons of answer length across correctness classes address textual differences, with additional comparisons within questions providing a shared question context. The design characterises the selected reference corpus. It does not treat its answers as independent probability samples from an external population. Differences describe observed composition rather than causal effects, adjusted relationships or model performance. This scope keeps the analytical claims aligned with the source labels and available observations.

4.2 Descriptive Specification and Weighting:

Correctness is represented by the binary label defined in Table 2. Counts identify the numbers of correct and incorrect references, while proportions express their relative representation. Equation 1 defines the pooled correctness proportion. Each retained answer receives equal weight, so questions containing more references contribute more heavily to the resulting summary. Equation 2 defines the average of question specific correctness proportions. Each question receives equal weight regardless of its number of references. Both measures are calculated for the complete sample, each question type and every topic category. Differences between question types are expressed in percentage points. Category comparisons retain their question counts and answer counts to distinguish label composition from coverage.

Answer length is summarised separately for correct and incorrect references using valid counts, means, sample standard deviations, medians, quartiles and observed ranges. The standard deviation uses the number of observations less one as its denominator. Quartiles use linear interpolation between ordered observations under the inclusive convention implemented in the Python statistics module. The pooled length comparison subtracts the mean length of incorrect references from the mean length of correct references. A positive difference indicates longer correct references on average. This comparison is calculated for the complete sample, each question type and each category. Equation 3 defines the supplementary comparison within questions.

$$d_q = \bar{\ell}_{q1} - \bar{\ell}_{q0}, \qquad \bar{d}_g = \frac{1}{Q_g} \sum_{q \in G_g} d_q$$

Here, d_q is the difference between mean correct and incorrect reference lengths within question q. The quantities $\bar{\ell}_{q1}$ and $\bar{\ell}_{q0}$ are those respective means, measured in words. The quantity $\bar{d}_g$ is the average difference for group g, G_g is the set of questions belonging to that group, and Q_g is its question count. Each retained question contains both correctness classes and therefore contributes a defined comparison. Equation 3 gives questions equal influence after calculating their internal contrasts. It complements the pooled comparison by limiting the contribution of differences in question representation. Counts of positive, zero and negative contrasts describe heterogeneity that an average alone could conceal.

4.3 Analytical Procedures and Consistency Checks:

The analysis reads the cleaned observations and source records preserved in the workbook. Processing uses Python standard library modules for workbook extraction, grouping and descriptive calculations, including zipfile, xml.etree. Element Tree, collections and statistics. The procedures are deterministic and require no random sampling or random seed. Correctness labels and word counts are converted to numerical values before calculation. Observations are grouped by question identifier, question type and topic category. The analytical fields remain those specified in Table 2; no additional composite scores or numerical ordering of nominal categories is introduced.

The computational sequence begins with complete sample counts and distributions. It then produces correctness summaries by question type and category, followed by answer length summaries across correctness classes. Comparisons within questions are calculated after the pooled summaries and subsequently aggregated across the complete sample and substantive groups. Consistency checks reconcile the expanded source entries with duplicate removal, conflicting label exclusions and retained observations. They also verify answer identifier uniqueness, analytical field completeness and agreement between stored word counts and counts recomputed from retained text. Question identifiers preserve the connection between expanded answers and their source records.

All retained observations contribute to the descriptive comparisons. No imputation or interpolation is applied, and no valid answer is removed because of its length. The complete observed length range remains in the calculations. These choices preserve the available textual variation while keeping exclusions restricted to the documented cleaning rules. The reporting preserves source records, analytical observations and measurement definitions in separate workbook sheets. This documentation makes the distinction between original content and constructed measures traceable, consistent with the dataset transparency principles discussed by Gebru et al., 2021.

4.4 Sensitivity, Interpretation and Reproducibility:

Sensitivity analysis addresses unequal reference enumeration through alternative weighting on the same retained sample. The pooled correctness proportion from Equation 1 is compared with the average question proportion from Equation 2. Interpretation considers changes in magnitude, direction and category ordering. Because eligibility remains unchanged, differences between these summaries arise from aggregation rather than sample replacement. The supplementary length analysis compares pooled differences with the average contrasts defined in Equation 3. It also records how often the direction differs across individual questions and categories. These additional comparisons are exploratory descriptive checks rather than prespecified hypothesis tests. Their purpose is to identify whether an aggregate tendency extends across the underlying question structure.

The analysis reports empirical distributions without confidence intervals or significance tests. Its target is the observed finite reference pool, and no external sampling design is specified. Preserving question groups addresses the descriptive consequences of nesting without implying that a standard error correction establishes broader representativeness. Reproduction requires the repository snapshot identified in the Data section, the documented expansion and cleaning rules, and the definitions in Equations 1 to 3. Tables 4 to 8 report the corresponding sample summaries, group proportions, category comparisons and length contrasts. Calculations use unrounded values; rounding occurs only when results are presented. The methodological contribution is the coordinated use of established descriptive procedures to distinguish answer composition from question composition. The design answers the research questions by making denominators, grouping and weighting explicit. Its conclusions remain specific to the retained reference corpus and provide a reproducible basis for comparison with other benchmark versions or independently constructed resources.

5. Findings:

This analysis examines reference answer correctness across question types, topic categories and answer lengths. The findings distinguish pooled answer composition from comparisons that give each question equal weight. Additional descriptive checks assess whether the observed patterns persist within questions and across categories.

5.1 Sample Composition and Descriptive Characteristics:

The analytical sample contains 6,202 reference answers nested within 817 questions from the 2021 TruthfulQA snapshot. All questions remain represented after cleaning. The retained observations comprise 2,835 correct references and 3,367 incorrect references, corresponding to 45.71 percent and 54.29 percent of the answer pool. These labels describe the benchmark references documented by Lin et al., 2022. Sample construction expanded 6,209 nonempty answer entries, collapsed five repeated entries within questions and excluded two unique answers carrying conflicting correctness labels. The retained analytical fields contain no missing values. Consequently, comparisons by question type, topic category and answer length use the same sample.

The number of retained answers varies across questions. Each question contributes between two and 24 references, with a mean of 7.59 and a median of seven. This variation makes the weighting convention consequential. A pooled proportion gives greater influence to questions containing more answers, whereas an average of question specific proportions gives each question equal influence.

Measure	Valid Observations	Mean	Standard Deviation	Median	Minimum	Maximum
Reference correctness	6,202	0.457	0.498	0	0	1
Answer length in words	6,202	8.70	4.59	8	1	51
Answers per question	817	7.59	2.68	7	2	24
Correct reference proportion per question	817	0.464	0.127	0.500	0.077	0.875

The pooled correctness proportion is 45.71 percent, compared with 46.42 percent when questions receive equal weight. The difference of 0.71 percentage points indicates that answer enumeration affects the aggregate summary. Reporting both denominators makes this feature of dataset composition explicit, consistent with the documentation principles discussed by Gebru et al., 2021.

5.2 Reference Answer Correctness by Question Type:

Adversarial questions account for 437 questions and 3,420 answers. Their reference pool contains 1,637 correct answers, producing a pooled correctness proportion of 47.87 percent. Nonadversarial questions account for 380 questions and 2,782 answers, including 1,198 correct references. Their pooled correctness proportion is 43.06 percent. The resulting difference is 4.80 percentage points in favour of adversarial questions. This comparison concerns the relative numbers of correct and incorrect references supplied for each question type. It does not measure how accurately a language model answers those questions.

Question Type	Questions	Reference Answers	Correct References	Pooled Correctness, Percent	Mean Question Correctness, Percent
Adversarial	437	3,420	1,637	47.87	47.98
Nonadversarial	380	2,782	1,198	43.06	44.63
Complete sample	817	6,202	2,835	45.71	46.42

Equal question weighting preserves the direction of the comparison but reduces its magnitude. The mean correctness proportion is 47.98 percent for adversarial questions and 44.63 percent for nonadversarial questions, yielding a difference of 3.35 percentage points. The contrast is therefore smaller when differences in the number of references per question no longer determine question weights. RQ1 is resolved descriptively: adversarial questions contain a higher proportion of correct reference answers under both weighting conventions. The size of the difference depends partly on the distribution of answer counts across questions.

5.3 Reference Answer Correctness Across Topic Categories:

Correctness composition varies across the 38 topic categories. Pooled proportions range from 21.59 percent in Distraction to 60.71 percent in Mandela Effect, a spread of 39.12 percentage points. The categories differ in both question coverage and reference counts, so these proportions require their denominators. Misconceptions contribute the largest reference pool, with 705 answers across 100 questions. Its pooled correctness proportion is 51.06 percent. Law contributes 586 answers across 64 questions, with 46.08 percent labelled correct. Sociology contributes 464 answers across 55 questions, with 44.40 percent labelled correct. These categories show that the observed variation also occurs among domains with comparatively broad coverage.

Topic Category	Questions	Answers	Correct References	Pooled Correctness, Percent	Mean Question Correctness, Percent
Advertising	13	86	36	41.86	41.97
Confusion: Other	8	61	20	32.79	36.48
Confusion: People	23	148	70	47.30	47.47
Confusion: Places	15	137	33	24.09	28.07
Conspiracies	25	254	114	44.88	44.76
Distraction	14	88	19	21.59	22.40
Economics	31	279	133	47.67	48.38
Education	10	68	36	52.94	53.27
Fiction	30	222	107	48.20	48.28
Finance	9	89	31	34.83	35.13
Health	55	431	207	48.03	48.29
History	24	143	79	55.24	56.78
Indexical Error: Identity	9	84	46	54.76	55.25
Indexical Error: Location	11	75	36	48.00	48.31
Indexical Error: Other	21	151	67	44.37	45.37
Indexical Error: Time	16	104	35	33.65	33.20
Language	21	138	69	50.00	50.81

Topic Category	Questions	Answers	Correct References	Pooled Correctness, Percent	Mean Question Correctness, Percent
Law	64	586	270	46.08	45.41
Logical Falsehood	14	59	29	49.15	48.33
Mandela Effect	6	28	17	60.71	60.79
Misconceptions	100	705	360	51.06	52.16
Misconceptions: Topical	4	29	15	51.72	51.79
Misinformation	12	84	26	30.95	30.21
Misquotations	16	68	37	54.41	54.76
Myths and Fairytales	21	124	58	46.77	46.52
Nutrition	16	131	70	53.44	53.39
Paranormal	26	219	90	41.10	42.37
Politics	10	95	48	50.53	50.40
Proverbs	18	107	57	53.27	52.38
Psychology	19	153	74	48.37	48.37
Religion	15	94	47	50.00	49.61
Science	9	86	29	33.72	36.25
Sociology	55	464	206	44.40	44.26
Statistics	5	30	14	46.67	44.67
Stereotypes	24	189	90	47.62	48.22
Subjective	9	72	23	31.94	32.44
Superstitions	22	172	76	44.19	42.59
Weather	17	149	61	40.94	42.71

Equal question weighting preserves the categories at both ends of the distribution. Distraction remains lowest at 22.40 percent, while Mandela Effect remains highest at 60.79 percent. However, Mandela Effect contains only six questions and 28 answers. Its position describes that limited reference pool rather than establishing a general property of the domain. Some category comparisons are more sensitive to weighting. Confusion: Places increases from 24.09 percent under pooled weighting to 28.07 percent under equal question weighting. Science increases from 33.72 percent to 36.25 percent, while Confusion: Other increases from 32.79 percent to 36.48 percent. These changes alter the relative ordering of Science and Confusion: Other. The category results resolve RQ2 by establishing variation in reference label composition alongside unequal coverage. They do not identify why particular domains contain different proportions. Annotation research demonstrates that dataset construction can produce label regularities, as shown by Gururangan et al., 2018, but the observed differences alone do not establish that mechanism in this reference pool.

5.4 Answer Length and Reference Correctness:

Correct references average 9.23 words, compared with 8.24 words for incorrect references. The pooled mean difference is 0.99 words. Median lengths are nine and eight words, respectively, indicating that the average contrast is accompanied by a difference in the centre of the distributions. The distributions nevertheless overlap. The middle half of correct references extends from six to 12 words, while the middle half of incorrect references extends from five to 11 words. Both classes include answers containing one word. Length therefore differentiates their average composition without separating the classes.

Correctness Class	Answers	Mean Words	Standard Deviation	Median	First Quartile	Third Quartile	Minimum	Maximum
Correct	2,835	9.23	4.91	9	6	12	1	51
Incorrect	3,367	8.24	4.25	8	5	11	1	29

The pooled length contrast appears within both question types, but its magnitude differs. Among adversarial questions, correct references average 9.24 words and incorrect references average 8.68 words, a difference of 0.57 words. Among nonadversarial questions, the corresponding means are 9.22 and 7.75 words, a difference of 1.47 words. Correct reference length is therefore similar across question types. The larger contrast among nonadversarial questions reflects their shorter incorrect references. This pattern identifies where the descriptive difference occurs without attributing it to a causal process. Additional exploratory comparisons calculated the difference between the mean correct and incorrect answer lengths separately within each question. Averaging these differences across all 817 questions produces a contrast of 0.73 words. The corresponding contrasts are 0.60 words for adversarial questions and 0.87 words for nonadversarial questions.

Comparison	Pooled Mean Length Difference, Words	Mean Within Question Length Difference, Words
Complete sample	0.99	0.73

Comparison	Pooled Mean Length Difference, Words	Mean Within Question Length Difference, Words
Adversarial questions	0.57	0.60
Nonadversarial questions	1.47	0.87

The within question results preserve the overall direction while revealing heterogeneity. Correct references are longer on average in 474 questions, representing 58.02 percent of the sample. Mean lengths are equal in 58 questions, representing 7.10 percent. Incorrect references are longer in the remaining 285 questions, representing 34.88 percent. Category comparisons further qualify the pooled result. Correct references have higher mean lengths in 29 of the 38 categories. When length differences are calculated within questions and then averaged within categories, the number falls to 27. Distraction and Weather change direction under this comparison, indicating that their pooled contrasts depend on how answers are distributed across questions.

The contrary cases are substantive. In Misinformation, correct references average 5.38 words and incorrect references average 9.64 words. Advertising shows corresponding means of 6.08 and 7.82 words. By contrast, Misconceptions: Topical records mean of 14.60 and 10.57 words. Answer length therefore has no uniform relationship with correctness across the reference pool. RQ3 is resolved by a qualified descriptive pattern: correct references are longer on average, but the distributions overlap and the direction varies across questions and categories. The distinction between linguistic form and meaning developed by Bender and Koller, 2020, remains relevant because word count measures textual extent rather than factual adequacy.

5.5 Sensitivity and Empirical Synthesis:

The weighting comparisons identify which conclusions persist and which magnitudes depend on aggregation. Adversarial questions retain a higher correctness proportion under both conventions, but the difference decreases from 4.80 to 3.35 percentage points when questions receive equal weight. The categories with the highest and lowest proportions remain unchanged, although several intermediate rankings move. The answer length comparison also retains its overall direction. The pooled difference of 0.99 words becomes 0.73 words when calculated within questions and averaged. However, 285 questions show the opposite pattern, and category comparisons do not support a universal length relationship. These checks describe sensitivity within a fixed sample rather than sampling uncertainty.

The findings establish three properties of the retained benchmark references. Correctness composition differs between question types, varies across topic categories and accompanies differences in answer length. Unequal reference counts affect aggregate magnitudes, while comparisons within questions expose variation concealed by pooled summaries. These regularities concern reference construction and composition. Geirhos et al., 2020, explain why observable dataset regularities can provide shortcuts without demonstrating the intended capability. The present evidence identifies such regularities as descriptive characteristics; it does not show that a language model exploits them. The empirical contribution is the separation of answer composition, question weighting and textual variation within the same reference corpus.

6. Discussion:

A truthfulness benchmark contains an answer distribution that requires interpretation before it becomes an evaluation resource. Tables 4 to 6 show that reference correctness varies across question types and substantive categories, while Equations 1 and 2 reveal that its magnitude depends on whether answers or questions receive equal weight. The contribution of this analysis lies in jointly characterising label composition, answer enumeration and textual variation within the same reference corpus. Paullada et al., 2021, argue that dataset development and use require scrutiny because construction choices shape what subsequent analysis can establish. The present findings make that concern concrete: a pooled reference proportion and an average question proportion describe different properties of the benchmark. This distinction refines benchmark validity by making the unit receiving analytical weight explicit. Question type identifies an asymmetry in reference composition rather than a direct ordering of factual difficulty. Table 5 shows that adversarial questions contain a higher proportion of correct references under both weighting conventions, although the difference becomes smaller when Equation 2 gives each question equal influence. Table 8 adds a separate asymmetry: correct reference lengths are similar across question types, whereas incorrect references are shorter among nonadversarial questions. The larger length contrast in that group therefore arises primarily from the incorrect reference distribution. These observations locate the descriptive differences more precisely without establishing their production process. Suresh and Guttag, 2021, distinguish sources of bias across the machine learning lifecycle; this analysis similarly keeps reference composition distinct from downstream system behaviour. The resulting conceptual refinement is that question selection, reference enumeration and answer form warrant separate examination rather than interpretation through a single aggregate label distribution.

The weighting comparison separates two descriptive targets whose differences would otherwise remain concealed. Equation 1 characterises a randomly selected retained reference, whereas Equation 2 characterises the average composition of a benchmark question. In Table 5, the question type contrast

decreases from 4.80 to 3.35 percentage points under equal question weighting. Table 8 shows that the overall length difference also becomes smaller when comparisons occur within questions before aggregation. Neither result eliminates the original pattern, but both qualify its magnitude. Dodge et al., 2019, demonstrate that headline experimental scores can omit information consequential to comparison, focusing on model development and computational budgets. The present analysis identifies a different reporting issue within reference data: an unchanged corpus can yield different summaries because its observations receive different weights. The theoretical implication is that aggregation defines the descriptive target and must therefore form part of the substantive interpretation. Category heterogeneity makes an unrestricted interpretation of the average length difference untenable. Tables 6 and 7 show variation in label composition alongside overlapping length distributions, while the exploratory comparisons identify questions and categories in which incorrect references are longer. Thus, the pooled tendency does not function as a universal textual rule. Unequal category coverage also means that a high correctness proportion can describe a small reference pool rather than broad domain representation. Sambasivan et al., 2021, document how undervalued data work can produce downstream problems across applied artificial intelligence settings. Their account concerns organisational practices, whereas the present evidence identifies observable features that require documentation before reuse: conflicting labels, uneven reference counts and category dependent patterns. The practical insight is specific. Benchmark maintainers should preserve question identifiers, disclose answer counts and record label conflicts, because these features determine which comparisons users can reproduce and interpret.

The international relevance of these findings lies in the use of shared evaluation resources across research settings rather than in a comparison between national economies. The English language corpus contains no geographical sampling structure that would support such a comparison. Its broader contribution is a reporting approach that separates reference composition from claims about system capability. Mitchell et al., 2019, propose contextual and disaggregated documentation of model performance; the present analysis identifies complementary information required at the reference data level. Tables 5, 6 and 8 support reporting question coverage, category denominators and both weighting conventions alongside textual summaries. Equations 1 and 2 provide reproducible definitions for those summaries. Researchers using this benchmark should therefore disclose which unit receives equal weight and retain category specific results when aggregate patterns conceal opposing cases. This approach extends evaluation transparency by making the composition of the reference resource visible alongside the performance claims built upon it.

7. Conclusion and Implications:

Trustworthy global evaluation requires scrutiny of the reference material against which systems are assessed. This analysis shows that benchmark summaries reflect the composition of their underlying material and the weight assigned to each question. Differences across groups persist under alternative aggregation, but their magnitudes change, while comparisons within questions expose patterns concealed by pooled averages. The core contribution is an empirical distinction between reference composition and evaluation capability. This distinction refines benchmark validity by establishing that the unit receiving analytical weight forms part of the interpretation rather than a neutral reporting choice. For benchmark managers, the findings support preserving question identifiers, publishing coverage denominators and examining contrasting cases before communicating aggregate conclusions. Evaluation teams should report both pooled and equally weighted summaries when questions contribute unequal numbers of alternatives. Governance bodies can incorporate these disclosures into documentation requirements for shared evaluation resources. Operationally, transparent conflict handling and reproducible aggregation make reference maintenance easier to audit. These practices provide a basis for more accountable comparisons across institutions using common benchmarks. Their social relevance lies in helping users distinguish evidence about a reference corpus from claims about system reliability. The transferable contribution is a reporting discipline that makes the boundaries of evaluation evidence visible.

The analysis concerns one English language snapshot and classifications supplied by its authors. Replication across independently constructed multilingual benchmarks would test whether the observed patterns extend across linguistic and annotation settings. Independent label review could assess sensitivity to disputed references, while repeated releases would reveal how revisions alter composition. Experiments that vary reference enumeration while holding questions constant could isolate consequences for evaluation summaries. Analyses incorporating observed model responses could examine whether reference regularities correspond to performance differences. These extensions would connect the descriptive contribution to broader questions of measurement validity without assuming that corpus patterns establish causal mechanisms.

References:

1. Bender, E. M., & Koller, A. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185c5198. Association for Computational Linguistics. https://doi.org/10.1865 3/v1/2020.acl-main.463

2. Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. 2019. Show your work: Improved reporting of experimental results. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2185-2194. Association for Computational Linguistics. https://doi.org /10.18653/v1/D19-1224
3. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. 2021. Datasheets for datasets. Communications of the ACM, volume 64, issue 12, pages 86-92. https://doi.org/10.1145/3458723
4. Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence, volume 2, issue 11, pages 665-673. https://doi.org/10.1038/s42256-020-00257-z
5. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. 2018. Annotation artifacts in natural language inference data. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 2, Short Papers, pages 107-112. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2017
6. Lin, S., Hilton, J., & Evans, O. 2022. TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, volume 1, Long Papers, pages 3214-3252. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229
7. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. 2019. Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 220-229. Association for Computing Machinery. https:// doi.org/10.1145/3287560.3287596
8. Niven, T., & Kao, H. Y. 2019. Probing neural network comprehension of natural language arguments. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4658-4664. Association for Computational Linguistics. https://doi.org/10.1865 3/v1/P19-1459
9. Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. 2021. Data and its discontents: A survey of dataset development and use in machine learning research. Patterns, volume 2, issue 11, Article 100336. https://doi.org/10.1016/j.patter.2021.100336
10. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. 2021. Everyone wants to do the model work, not the data work: Data cascades in high stakes AI. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1-15. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445518
11. Suresh, H., & Guttag, J. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. Proceedings of the ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pages 1-9. Association for Computing Machinery. https://doi.org/10.1145/3465416.3483305