International Journal of Advanced Trends in Engineering and Technology (IJATET)
International Peer Reviewed - Refereed Research Journal, Website: www.dvpublication.com
Impact Factor: 5.965, ISSN (Online): 2456 - 4664, Volume 11, Issue 2, July - December, 2026

MODEL CONFIGURATION AND TASK DESIGN ASSOCIATIONS WITH HALLUCINATION TEST ACCURACY IN MEDICAL LANGUAGE MODEL BENCHMARKS

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, "Model Configuration and Task Design Associations with Hallucination Test Accuracy in Medical Language Model Benchmarks", International Journal of Advanced Trends in Engineering and Technology, Volume 11, Issue 2, July - December, Page Number 5-13, 2026.
Copy Right: © DV Publication, 2026 (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.
Type of Review: Peer Reviewed as per |C|O|P|E| Guidance.
Disclaimer: The scholarly papers reviewed and published by DV Publication, India, reflect the views and opinions of their respective authors and do not necessarily represent the views or opinions of DV Publication. The publisher disclaims any responsibility for any harm, loss, or damage resulting from the use of the published content by any party.
DOI: https://doi.org/10.5281/zenodo.24041087

Abstract:

This analysis examines whether medical language model rankings remain consistent across configurations and evaluation tasks, informing international debates on reliable model selection. Published benchmark results from 2023 provide 70 observations covering ten configurations and seven tasks. Descriptive comparisons examine model families, parameter sizes within Llama 2, and 35 matched differences between base and tuned variants. Family rankings vary across tasks. Average base Llama 2 accuracy increases from 18.95 to 35.59 percent between seven and seventy billion parameters, whereas tuned accuracy peaks at thirteen billion. Across matched comparisons, tuning accompanies a mean difference of −9.30 percentage points, but the median difference is zero. Reasoning comparisons average −26.96 percentage points, while memory comparisons average 3.94 percentage points, exposing contrasting performance patterns concealed by aggregation. These descriptive results refine capability and task alignment by demonstrating that configuration advantages depend on evaluation demands rather than a uniform hierarchy. The contribution is a transparent comparison framework that distinguishes improvement frequency from magnitude. Institutions should require task specific evidence when selecting configurations and avoid interpreting aggregate benchmark superiority as demonstrated clinical reliability.

Key Words: Medical Language Models; Hallucination Evaluation; Instruction Tuning; Benchmark Comparison; Descriptive Analysis

1. Introduction:

Selecting a medical language model requires knowing which capabilities it performs reliably, not simply which configuration achieves the highest overall score. A model may reject fabricated questions successfully while struggling to retrieve bibliographic information or evaluate a suggested answer. These differences matter when institutions compare model families, parameter sizes and conversational variants for applications with distinct information requirements. Hager et al., 2024, demonstrate that performance on medical examinations does not establish competence across information gathering, instruction following and clinical workflows. The substantive problem is therefore one of conditional performance: an advantage established under one evaluation demand may not extend to another. Medical hallucination benchmarks provide a structured basis for examining this problem, but their interpretation depends on what each task measures. Pal et al., 2023, introduce Med HALT to evaluate reasoning and memory related performance through false confidence, fabricated question, answer selection and bibliographic retrieval tasks. Their published results distinguish model configurations and identify differences associated with instruction tuning. However, task accuracy remains an evaluation outcome rather than a direct measure of clinical safety. Farquhar et al., 2024, distinguish confabulations from other incorrect outputs, showing why different sources of error require careful interpretation. This analysis consequently retains the published accuracy measure and its task identity rather than treating every incorrect response as an equivalent clinical hallucination.

Model family, parameter size and instruction tuning represent complementary dimensions of configuration. Family membership captures a package of development choices, while parameter size describes nominal representational capacity. Greater capacity may support stronger performance, but parameter count alone does not establish that a model will reject an unsupported premise or retrieve an exact identifier. Instruction tuning introduces another potential distinction by adapting response behaviour to conversational or instructional demands. Such adaptation may accompany improvements on some tasks and deterioration on others. These relationships motivate a capability and task alignment perspective in which configuration advantages are evaluated under specified demands. The framework concerns observed associations; the available configuration labels do not separately identify the training processes responsible for them. Aggregation creates a further interpretive challenge. An overall accuracy average can conceal task rankings, offsetting improvements and declines, or a small number of large differences. Agrawal et al., 2025, explain how discrepancies among data, tasks and metrics can produce misleading conclusions about medical language model performance. Mitchell et al., 2019, similarly advocate documenting model performance under the conditions relevant to intended use. The closest benchmark research already establishes substantial variation across configurations and tasks. The contribution pursued here is narrower: a secondary analysis that brings family comparisons, size patterns and matched tuning differences into a common descriptive framework. It examines whether conclusions based on aggregate performance remain informative when task identity and configuration matching are preserved.

This study aims to characterize how published hallucination test accuracy varies across model configurations within common medical benchmark tasks. The central research question asks how model family, parameter size and instruction tuning

relate to accuracy when comparisons retain the evaluation task. Three objectives operationalize this question: compare observed families within common tasks and tuning categories, examine parameter size patterns within Llama 2, and calculate accuracy differences between base and tuned configurations matched by family, size and task. The dataset contains 70 published observations covering ten configurations and seven tasks from the 2023 benchmark publication cohort. Descriptive summaries and 35 matched tuning comparisons preserve the structure of the available evidence without treating repeated task evaluations as independent model samples. The findings qualify broad configuration rankings. Family advantages change across tasks, increasing parameter size accompanies higher average base Llama 2 accuracy without a corresponding ordering among tuned variants, and matched tuning differences vary between reasoning and memory evaluations. The analytical contribution is to show how aggregate comparisons can conceal these conditional patterns, including negative mean differences alongside positive medians within some configuration pairs. Practically, the results support comparing the specific configuration and capabilities required for an intended application rather than relying on family reputation, parameter count or tuning status alone. The transferable insight concerns evaluation design and reporting across institutions; it does not establish performance across countries, languages or clinical deployments.

2. Theory, Conceptual Framework and Research Questions Development:

This analysis adopts a capability and task alignment perspective to explain variation in medical language model performance. Generated answers depend on the interaction between learned representations, model adaptation and the demands imposed by an evaluation task. Statistical fluency therefore does not guarantee factual correctness. Bender et al., 2021, distinguish the reproduction of linguistic patterns from grounded meaning, providing a theoretical basis for examining accuracy separately from the apparent coherence of generated responses. Here, model configuration comprises observed family membership, parameter size and instruction tuning status, while the outcome is published hallucination test accuracy. These measures support comparisons of benchmark performance without treating model configuration as a directly observed cognitive mechanism. The conceptual framework connects these configuration dimensions with accuracy within seven benchmark tasks. Accuracy denotes the percentage of responses classified as correct under each task's evaluation procedure. It does not measure a uniform clinical hallucination rate. Farquhar et al., 2024, distinguish confabulations from other sources of incorrect output, demonstrating why different error mechanisms should not be collapsed into a single interpretation. Accordingly, this analysis preserves task identity throughout the comparisons. The framework represents associations between observed configuration characteristics and observed accuracy, rather than causal pathways or latent constructs.

Model family captures differences among the Llama 2, Falcon and MPT configurations represented in the dataset. Family membership identifies a package of development choices rather than an isolated architectural treatment. Differences in learned representations and adaptation can influence whether a model retrieves relevant information, follows task instructions or generates a plausible but incorrect answer. However, these underlying processes are not separately measured. Thirunavukarasu et al., 2023, describe medical language model performance as dependent on development and evaluation conditions, supporting an interpretation that remains specific to the task and configuration being compared. The theoretical expectation is therefore heterogeneity across families, without a defensible universal ranking. The family comparison uses accuracy distributions within common tasks and reports base and tuned configurations separately. Because the families are represented by different parameter sizes, any observed family difference remains a configuration difference rather than an independent architectural effect. Research question 1 therefore asks: How does published hallucination test accuracy vary across the observed model families within the same benchmark tasks and tuning categories?

Parameter size represents the number of model parameters, expressed in billions. Greater representational capacity may support the encoding and retrieval of medical knowledge, creating an expectation of higher accuracy. Singhal et al., 2023, report improvements in medical question answering associated with model scale and adaptation. This evidence motivates examining size differences, but it does not establish that additional parameters improve every hallucination task. Larger capacity can strengthen learned associations without ensuring that the model rejects an unsupported premise or retrieves an exact bibliographic identifier. A consistent positive pattern therefore remains an empirical possibility rather than an assumption. The dataset permits a focused size comparison within Llama 2 because it contains base and chat configurations at 7, 13 and 70 billion parameters. Holding family, task and tuning category constant reduces the interpretive ambiguity of comparisons across unrelated configurations. The analysis describes ordered accuracy patterns across these sizes without imposing linearity or attributing differences solely to parameter count. Research question 2 therefore asks: Within Llama 2, how does published hallucination test accuracy vary across parameter sizes when benchmark task and tuning category remain the same?

Instruction tuning represents adaptation intended to make generated responses conform more closely to instructions or conversational expectations. Improved instruction following could increase accuracy when success depends on interpreting the requested response format. A competing mechanism arises when adaptation encourages answer production despite insufficient evidence or changes how the model responds to misleading questions. Pal et al., 2023, distinguish reasoning tests from memory tests and report configuration differences across these tasks. Their evidence supports examining tuning separately within each task rather than assuming that conversational adaptation produces a uniform factual advantage. The available configurations support 35 comparisons between base and tuned variants matched by family, parameter size and benchmark task. Each difference expresses tuned accuracy minus base accuracy in percentage points. Positive, negative and zero differences are retained because the direction may vary with task demands. Research question 3 therefore asks: Within matched family, size and task combinations, how does published hallucination test accuracy differ between base and instruction tuned configurations?

These research questions organize a descriptive analysis of 70 published model and task observations. Family comparisons characterize configuration differences, size comparisons examine variation within Llama 2, and matched comparisons describe tuning differences. The observations remain repeated evaluations of ten models rather than independent model samples. The analysis therefore uses task specific accuracy patterns and matched percentage point differences to resolve the research questions, without assigning causal meaning to the observed associations.

3. Data:

This analysis documents the provenance, selection and measurement of published medical language model benchmark results. The resulting dataset supports descriptive comparisons of model configurations within common evaluation tasks.

3.1 Data Sources, Population, and Sample Construction:

The empirical source is the published Med HALT benchmark reported by Pal et al., 2023, in the proceedings of the Association for Computational Linguistics. Task accuracy values were extracted from Table 2 on page 320 and Table 3 on page 321 of the published proceedings version. The analytical workbook uses published aggregate results rather than repository question records or newly generated model responses. The publication carries a Creative Commons Attribution 4.0 licence. The unit of analysis is one model configuration evaluated on one benchmark task. The dataset contains a single publication cohort from 2023 within the temporal eligibility boundary of 2010 to 2025. It forms a balanced configuration by task matrix, with ten configurations evaluated on seven tasks. It is not an annual panel. Accuracy values are evaluation percentages rather than flows, stocks or period averages. The empirical setting is medical reasoning and bibliographic information retrieval, with geographical coverage inherited from the benchmark rather than represented by country observations. The target population comprises medical language model configurations evaluated through hallucination related benchmark tasks. The accessible sampling frame is narrower: the twelve configurations reported across the seven tasks in the two source tables. Selection retained configurations with identifiable model family, parameter size and tuning status. GPT 3.5 and Text Davinci were excluded because comparable parameter sizes were unavailable. Published average columns were omitted because they summarize existing task results rather than provide additional observations. The selection rules are documented in the workbook. No dated preregistration or independent record establishes that they preceded inspection of the published results. The analytical sample therefore represents a documented secondary selection rather than a prespecified probability sample. Separating population boundaries, source composition and permitted uses follows the documentation logic described by Gebru et al., 2021.

Table 1: Sample construction

Stage	Model Configurations	Model and Task Observations	Reason or Rule
Published sampling frame	12	84	Twelve configurations evaluated on seven tasks
Exclusion of GPT 3.5	1 removed	7 removed	Comparable parameter size unavailable
Exclusion of Text Davinci	1 removed	7 removed	Comparable parameter size unavailable
Consolidated analytical sample	10	70	Required configuration characteristics and accuracy available
Retained descriptive sample	10	70	No additional missing value exclusions
Matched tuning subset	10	70 contributing observations	Five base and tuned configuration pairs produce 35 task specific differences

The retained sample contains six Llama 2 configurations, two Falcon configurations and two MPT configurations. These contribute 42, 14 and 14 observations, respectively. Five configurations are base models and five are chat or instruct variants. The matched subset uses the same observations and introduces no additional model evaluations.

3.2 Data Integration, Cleaning, and Quality Assurance:

The two published tables were consolidated vertically into a common schema. Each record contains a source model label, harmonized family identifier, parameter size, tuning indicator, task identifier, accuracy value, publication cohort and source location. Model family labels were standardized to Llama 2, Falcon and MPT while the original model labels were retained for traceability. The combination of model configuration and task defines the record key. Verification identified 70 unique configuration and task keys, with no duplicate keys or missing cells across the eleven recorded fields. Each configuration has seven task observations. Repeated configurations across tasks are intentional measurements, not duplicate records. No external dataset merge was performed, so unmatched records and merge induced observation increases do not arise. All retained accuracy values fall between zero and 100 percent. Zero values were preserved as published evaluation results rather than substituted for missing information. Values near either boundary were also retained. No winsorization, trimming, interpolation, imputation or rescaling was applied. Parameter sizes remain expressed in billions, and accuracy remains expressed in percent. The source publication is the authoritative version for numerical values. The workbook records an administrative extraction date separately from the 2023 publication cohort; that date does not identify when the benchmark responses were generated. Exact evaluation dates are not available in the analytical dataset. Consequently, the observations describe the published configurations and evaluation conditions rather than changes through calendar time. The workbook preserves source locations, selection rules, variable definitions and matched comparison formulas. These features provide a traceable route from the published tables to the analytical records, consistent with the computational documentation principles discussed by Sandve et al., 2013. No alternative eligibility analysis or independent replication of the original benchmark was implemented.

3.3 Variable Measurement, Construction, and Validation:

Model configuration is represented through three observed characteristics: family membership, parameter size and instruction tuning status. These characteristics are analyzed separately rather than combined into a composite index. Family membership identifies configuration groups, parameter size records nominal model capacity, and tuning status distinguishes base configurations from chat or instruct variants. The outcome is the task accuracy reported in the source publication. Higher values indicate greater success under the corresponding task evaluation procedure. Accuracy is not converted into a clinical hallucination rate because incorrect responses can arise through different mechanisms. Farquhar et al., 2024, distinguish confabulations from other incorrect outputs, supporting the preservation of the benchmark's task specific interpretation.

Table 2: Variable dictionary

Variable or Indicator	Analytical Role	Definition and Construction	Source and Identifier	Unit and Frequency	Transformation and Direction	Missing Value Rule	Coverage and Limitation
Task accuracy	Outcome	Published percentage of correct responses for each configuration and task	Published Tables 2 and 3	Percent; one evaluation cohort	Original values retained; higher indicates greater accuracy	No missing values among retained records	70 observations; task evaluation procedures differ
Model family	Configuration characteristic	Family mapped from the published model label	Source table row labels	Categorical; one evaluation cohort	Llama 2, Falcon or MPT; no numerical ordering	Complete among retained records	Three families with unequal configuration counts
Parameter size	Configuration characteristic	Nominal parameter count identified from the model label	Source table row labels	Billions of parameters; one evaluation cohort	Original size retained; higher indicates greater nominal capacity	Configurations without comparable sizes excluded	Values of 7, 13, 40 and 70; family and size partly overlap
Instruction tuning status	Configuration characteristic	Base equals zero; chat or instruct equals one	Source table row labels	Binary; one evaluation cohort	One identifies an adapted variant	Complete among retained records	Different adaptation procedures share one category
Benchmark task	Comparison stratum	False confidence, fake question, none of the above, and four bibliographic retrieval tasks	Source table column headings	Seven categories; one evaluation cohort	No ordered transformation	Complete for every retained configuration	Common task labels do not establish identical error mechanisms
Test group	Descriptive grouping	Reasoning or memory classification inherited from the source table	Table 2 or Table 3	Two categories; one evaluation cohort	No ordered transformation	Complete among retained records	Three reasoning tasks and four memory tasks
Publication cohort	Provenance indicator	Year of the published benchmark results	Proceedings version	Year; single cohort	Constant at 2023	Complete among retained records	Does not establish exact response generation dates
Matched tuning difference	Calculated comparison	Tuned accuracy minus base accuracy within the same family, size and task	Linked workbook accuracy cells	Percentage points; one evaluation cohort	Positive values favour the tuned configuration	Complete for 35 matched comparisons	Describes configuration differences without isolating a causal tuning effect

The matched tuning difference is calculated as

$$D_{gst} = A^{T}_{gst} - A^{B}_{gst}$$

Here, D_{gst} denotes the accuracy difference for family g, parameter size s and task t. A^{T}_{gst} denotes tuned accuracy, and A^{B}_{gst} denotes base accuracy. Equation 1 produces percentage point differences, not percentage changes. Each of the five family and size pairs contributes seven differences.

Table 3: Verified descriptive statistics

Measure	Valid Observations	Mean	Sample SD	Median	Minimum	Maximum
Task accuracy, percent	70	25.50	27.18	19.27	0.00	99.89
Parameter size, billions	70	27.40	24.73	13.00	7.00	70.00
Matched tuning difference, percentage points	35	−9.30	26.77	0.00	−91.77	33.62

The parameter summary describes the observation matrix, in which each configuration appears seven times. The pooled accuracy summary gives equal weight to configuration and task cells; it is not accuracy pooled across underlying questions. The standard deviation describes dispersion among published values rather than uncertainty around individual benchmark estimates.

3.4 Analytical Framework, Identification, and Reproducibility:

The implemented framework is descriptive. It compares published accuracy across model families within common tasks, examines size patterns within Llama 2, and calculates matched differences between base and tuned variants. The size comparison uses the 7, 13 and 70 billion parameter configurations separately within base and chat categories. The tuning comparison holds family, nominal size and task constant through the matching structure in Equation 1. The family comparison describes the configurations available in the benchmark. It does not isolate architecture because family membership overlaps with parameter size and other development choices. Likewise, matched tuning differences describe published variants without establishing that tuning alone generated the differences. The analysis introduces no regression estimator, fixed effects, interactions, significance tests or causal identification claim. The 70 records are repeated evaluations of ten configurations. They are therefore summarized as a structured benchmark matrix rather than treated as 70 independent models. Task specific comparisons preserve differences in evaluation demands, while the pooled summaries describe the composition of the available matrix. No uncertainty estimates are reconstructed from percentages because the workbook does not contain the underlying response records needed to reproduce them. The workbook contains the consolidated dataset, source identifiers, variable definitions, selection documentation and linked formulas for all 35 matched differences. These materials make the implemented calculations inspectable. Peng, 2011, identifies access to the data and computational steps underlying an analysis as central to computational reproducibility. Here, reproducibility concerns the secondary extraction and descriptive calculations; it does not imply reproduction of the original model evaluations.

4. Methods:

4.1 Research Design and Comparison Strategy:

This analysis uses a descriptive comparative design to examine published medical language model benchmark performance. The unit of analysis is one configuration evaluated on one task. The analytical targets are observed accuracy differences across families, accuracy patterns across parameter sizes within Llama 2, and matched differences between base and

tuned configurations. These comparisons use a single publication cohort rather than temporal variation. The design preserves task identity because the benchmark distinguishes reasoning evaluations from bibliographic retrieval evaluations. Pal et al., 2023, define these tasks through different prompts and evaluation requirements. Comparisons therefore concern performance under specified benchmark conditions rather than a common clinical error probability. Three comparison strategies address the research questions. Family comparisons hold task and tuning category constant. Size comparisons retain family membership and compare the three available Llama 2 sizes separately within base and chat categories. Tuning comparisons match configurations by family, nominal parameter size and task before calculating the accuracy difference. Matching improves comparability along observed characteristics but does not isolate a causal effect. Family membership incorporates multiple development choices, and nominally matched base and tuned variants may differ in adaptation procedures that are not separately recorded. Accordingly, all contrasts describe the retained configurations. The design does not estimate treatment effects, mediation pathways or statistical moderation.

4.2 Analytical Definitions and Summary Measures:

The sample construction rules and variable definitions follow Table 1 and Table 2. The retained matrix contains ten configurations evaluated on seven tasks, yielding 70 observations. Family comparisons use all retained configurations, while size comparisons use the six Llama 2 configurations. Five base and tuned configuration pairs contribute 35 matched task comparisons. For each comparison set, mean accuracy is calculated as

$$\overline{A}_C = \frac{\sum_{i \in C} A_i}{n_C}$$

Here, $\overline{A}_C$ denotes mean accuracy for comparison set C, A_i denotes the published accuracy percentage for observation i, and n_C denotes the number of observations in that set. Equation 2 applies to pooled summaries, individual tasks, family and tuning categories, and size specific task groups. Each observation receives equal weight within its comparison set. Consequently, a family and tuning mean contains three Llama 2 configurations but one Falcon or MPT configuration. Reasoning summaries average three tasks, memory summaries average four tasks, and overall configuration summaries average seven tasks. Overall means therefore reflect the benchmark's task composition rather than an application specific weighting scheme. Matched tuning differences follow Equation 1, defined as tuned accuracy minus base accuracy within the same family, size and task. Positive differences favour the tuned configuration, negative differences favour the base configuration, and zero denotes equality at the precision reported by the source. Differences are expressed in percentage points. No percentage change transformation is applied. The descriptive statistics comprise valid counts, arithmetic means, medians, minima, maxima and sample standard deviations. Standard deviations use the number of observations minus one as the denominator. Categorical characteristics are summarized through counts. Parameter size remains expressed in billions and is not combined with family membership or tuning status into an index.

4.3 Computation and Analytical Procedures:

Calculations were performed in Python 3.12.14 using the standard library. Workbook contents were read with the zipfile and xml.etree. Element Tree modules, and descriptive summaries were calculated with the statistics module. The calculations are deterministic and require no random seed, numerical optimization or convergence criterion. The analytical sequence begins with the cleaned configuration and task records. Records are grouped by task to describe differences in evaluation difficulty within the retained matrix. They are then grouped jointly by family, tuning category and task to answer the first research question. Within each group, the arithmetic mean is calculated using Equation 2. The second research question is examined by restricting the records to Llama 2 and ordering nominal parameter sizes as seven, thirteen and seventy billion. Accuracy is compared across these sizes within each task and tuning category. Separate reasoning, memory and overall averages provide complementary summaries. An increasing overall average is not interpreted as consistent improvement unless the individual task comparisons also follow that ordering. For the third research question, each tuned record is linked to its base counterpart through the joint family, size and task key. Equation 1 produces the 35 matched differences. These differences are summarized across all pairs, separately within reasoning and memory groups, and within each family and size pair. Counts of positive, negative and unchanged differences distinguish the frequency of improvement from its magnitude. No observations are removed during these calculations. Published zero and extreme accuracy values remain in the sample. No imputation, interpolation, winsorization or transformation of the outcome is applied. Calculations retain numerical precision until presentation, when reported percentages and percentage point differences are rounded to two decimal places.

4.4 Verification, Reproducibility and Interpretive Limits:

Verification checks assess whether the analytical structure matches the intended comparisons. The configuration and task key is unique across all 70 records, each configuration contributes seven tasks, and the required fields contain no missing values. Accuracy values are checked against the permissible interval from zero to 100 percent. Matching confirms one base counterpart for each tuned observation and produces five configuration pairs across seven tasks. The numerical summaries are generated from the same retained records. Family, size and tuning comparisons therefore differ through their grouping rules rather than through undocumented sample exclusions. Reporting means alongside medians and sign counts examines whether aggregate differences conceal uneven magnitudes or opposing directions. Separating reasoning and memory summaries examines how task composition affects the pooled interpretation; it does not constitute an independent replication. The observations are repeated evaluations of ten configurations, not 70 independent model samples. The analysis therefore reports descriptive magnitudes without standard errors, confidence intervals or significance tests. Published aggregate percentages do not provide the response records needed to reconstruct uncertainty or dependence among answers. Dispersion statistics describe variation within the observed matrix rather than sampling precision. The workbook preserves source locations, original configuration labels, measurement definitions and linked formulas for the matched comparisons. These records support tracing the calculations from published values to reported outputs. Sandve et al., 2013, emphasize preserving computational inputs and transformations, while Peng, 2011, distinguishes reproducible computation from independent replication. The reproducibility claim here concerns the secondary calculations; it does not imply reproduction of the original model evaluations. No preregistration is documented, and the descriptive grouping

choices are not represented as prespecified tests. The method applies established summaries and matching rules to answer bounded research questions. Its contribution lies in transparent comparison and interpretation rather than a new estimator or identification strategy.

5. Findings:

This analysis examines how published hallucination test accuracy varies across model families, parameter sizes and tuning categories. The findings address the three research questions through comparisons within common tasks and matched configurations. All reported differences describe the observed benchmark results and retain their task specific interpretation.

5.1 Sample Composition and Accuracy Distribution:

The analytical sample contains 70 observations covering ten model configurations and seven benchmark tasks in the 2023 publication cohort. Llama 2 contributes six configurations, while Falcon and MPT contribute two each. Every configuration has complete observations for all seven tasks. The comparisons therefore use the same underlying sample without losses from missing values. Accuracy varies substantially across configuration and task combinations. The pooled mean is 25.50 percent, the median is 19.27 percent, and the sample standard deviation is 27.18 percentage points. Observed values range from zero to 99.89 percent. The mean exceeds the median, indicating that relatively high accuracy values raise the pooled average. These summaries give equal weight to each configuration and task observation rather than weighting by the number of underlying benchmark questions.

Table 4: Accuracy across benchmark tasks

Benchmark Task	Configurations	Mean Accuracy, Percent
False confidence	10	8.69
Fake question	10	56.46
None of the above	10	45.95
PMID to title	10	10.39
Title to PubMed link	10	29.13
Abstract to PubMed link	10	25.26
PubMed link to title	10	2.62

Table 4 identifies task composition as essential to interpreting model performance. Mean accuracy reaches 56.46 percent for fake questions but falls to 2.62 percent for generating a title from a PubMed link. False confidence accuracy is also low at 8.69 percent. Consequently, a configuration's overall average can conceal substantial differences between rejecting problematic questions and retrieving bibliographic information.

5.2 Model Family Differences Within Common Tasks:

The first research question concerns differences across observed model families within the same task and tuning category. Table 5 presents these comparisons while preserving the distinction between base and tuned configurations.

Table 5: Mean accuracy by family, tuning category and task, percent

Benchmark Task	Llama 2 Base	Llama 2 Tuned	Falcon Base	Falcon Tuned	MPT Base	MPT Tuned
False confidence	14.79	7.24	18.66	1.11	0.85	0.17
Fake question	81.81	16.31	99.89	99.35	48.49	22.55
None of the above	73.80	26.50	58.72	55.69	19.88	24.34
PMID to title	0.22	0.73	40.46	40.46	20.08	0.04
Title to PubMed link	9.69	34.21	40.46	40.46	40.46	38.24
Abstract to PubMed link	7.39	23.01	40.46	40.44	40.03	40.46
PubMed link to title	7.91	0.78	0.06	0.08	0.00	0.00

Family rankings change across tasks. Among base configurations, Falcon has the highest false confidence accuracy at 18.66 percent and fake question accuracy at 99.89 percent. Llama 2 leads on the none of the above task, with an average of 73.80 percent compared with 58.72 percent for Falcon. For title to PubMed link retrieval, Falcon and MPT share the highest base accuracy at 40.46 percent. Tuning changes these comparisons. Llama 2 has the highest tuned false confidence accuracy at 7.24 percent, while Falcon leads the tuned fake question and none of the above tasks at 99.35 and 55.69 percent. MPT narrowly exceeds Falcon on abstract to PubMed link retrieval, with 40.46 versus 40.44 percent. That 0.02 percentage point difference is small relative to the larger differences elsewhere in the matrix. Research question 1 is therefore resolved descriptively: observed family differences depend on the benchmark task and tuning category. The evidence does not establish a uniform family ranking. Because family membership overlaps with parameter size, these comparisons describe the available configurations rather than isolate architecture.

5.3 Parameter Size Patterns Within Llama 2:

The second research question examines parameter size within Llama 2 while retaining common task and tuning categories. Table 6 summarizes accuracy separately for reasoning and memory tasks.

Table 6: Mean Llama 2 accuracy by parameter size and tuning category, percent

Parameters, Billions	Base Reasoning	Base Memory	Base All Tasks	Tuned Reasoning	Tuned Memory	Tuned All Tasks
7	42.89	1.00	18.95	17.83	10.93	13.89
13	55.18	9.88	29.29	20.95	20.07	20.45
70	72.33	8.04	35.59	11.26	13.05	12.28

Base accuracy increases across the three sizes when averaged over all tasks, rising from 18.95 percent at seven billion parameters to 35.59 percent at 70 billion. The reasoning average increases by 29.45 percentage points between these configurations. At the individual task level, base accuracy increases consistently with size for false confidence, fake questions, none of the above, title to PubMed link retrieval and abstract to PubMed link retrieval. Memory performance does not show the same ordering. The base memory average rises from 1.00 percent at seven billion parameters to 9.88 percent at 13 billion, then declines to 8.04 percent at 70 billion. PubMed link to title accuracy illustrates this discontinuity: the respective values are zero, 23.72 and 0.02 percent. The largest base configuration therefore does not dominate every retrieval task. Tuned configurations show a different size pattern. Overall accuracy peaks at 20.45 percent for the 13 billion parameter configuration, compared with 13.89 percent at seven billion and 12.28 percent at 70 billion. Tuned false confidence accuracy increases consistently with size, but fake question accuracy falls from 21.96 percent at seven billion parameters to 5.49 percent at 70 billion. Research question 2 is resolved through this distinction: increasing size accompanies higher average base accuracy, particularly on reasoning tasks, but the relationship does not extend consistently across memory tasks or tuned configurations. The observed pattern supports conditional comparisons rather than a general claim that larger configurations perform better.

5.4 Matched Differences Between Base and Tuned Configurations:

The third research question compares base and tuned accuracy within identical family, parameter size and task combinations. Across 35 matched comparisons, tuned accuracy is higher in 13 cases, lower in 17 and unchanged in five. The mean difference is −9.30 percentage points, while the median is zero. The contrast between these summaries shows that several large negative differences pull the mean below the typical matched difference.

Table 7: Matched tuning differences by configuration pair

Family	Parameters, Billions	Comparisons	Mean Difference, Percentage Points	Median Difference, Percentage Points	Higher	Lower	Unchanged
Llama 2	7	7	−5.06	0.00	2	3	2
Llama 2	13	7	−8.85	0.85	4	3	0
Llama 2	70	7	−23.31	0.59	4	3	0
Falcon	40	7	−3.02	−0.02	1	4	2
MPT	7	7	−6.28	−0.68	2	4	1
All pairs	Mixed	35	−9.30	0.00	13	17	5

Every configuration pair has a negative mean difference, but the direction is not consistent across its tasks. The 70 billion parameter Llama 2 pair records four positive differences and three negative differences, yet its mean difference is −23.31 percentage points. Its largest decline occurs on fake questions, where accuracy falls from 97.26 to 5.49 percent, a difference of −91.77 percentage points. Conversely, the largest positive difference is 33.62 percentage points for abstract to PubMed link retrieval in the 13 billion parameter Llama 2 pair. The task grouping exposes a systematic contrast within the observed matrix. Across 15 reasoning comparisons, 13 differences are negative and two are positive. The mean reasoning difference is −26.96 percentage points and the median is −25.94. Across 20 memory comparisons, 11 differences are positive, four are negative and five are unchanged. The mean memory difference is 3.94 percentage points, while the median is 0.23. Research question 3 is therefore resolved descriptively: tuned variants have lower accuracy on average across the full matrix, with declines concentrated in reasoning tasks and mixed improvements in memory tasks. The pooled negative difference does not imply deterioration on every task. Together, the findings establish that configuration comparisons require joint attention to family, size, tuning category and evaluation task.

6. Discussion:

Medical language model performance cannot be reduced to a single ranking. Tables 4 and 5 show that family comparisons change with the evaluation task and tuning category, while Table 6 shows that parameter size does not produce a consistent ordering across configurations. The contribution of this analysis is the joint descriptive interpretation of these conditional patterns: configuration advantages remain specific to the capability being evaluated. This sharpens the distinction between benchmark performance and general medical competence. Agrawal et al., 2025, identify discrepancies among evaluation data, tasks and metrics as sources of misleading performance claims. The present comparisons give that concern a concrete expression: averaging across tasks can conceal reversals in the relative performance of available configurations. The framework therefore treats capability as a profile across tasks rather than a scalar property of model family or size. The matched comparisons reveal that instruction tuning accompanies different performance patterns across reasoning and memory tasks. Equation 1 holds family, nominal parameter size and benchmark task constant when calculating each difference, making the comparison more specific than an unrestricted comparison of model averages. Table 7 shows a negative overall mean alongside a zero median, while the task group analysis locates the larger declines in reasoning evaluations and mixed improvements in memory evaluations. This asymmetry changes the interpretation of adaptation: an improvement in one evaluated capability does not establish improvement across the benchmark. It also directs attention to the distinction between producing an instruction compatible response and producing a correct response. Hager et al., 2024, demonstrate that instruction following, diagnostic performance and workflow compatibility require separate evaluation. Here, the evidence identifies where configuration differences occur, while leaving the underlying training processes outside the measured explanation.

Parameter size provides a second boundary to general performance claims. Table 6 shows increasing average accuracy across base Llama 2 sizes, but a different ordering among tuned configurations and no consistent improvement across memory tasks. The theoretical implication is conditional: greater nominal capacity accompanies stronger performance in some comparisons without establishing a universal capacity advantage. The contrast between reasoning and memory summaries also explains why pooled results depend on the tasks included. Because the overall mean assigns equal weight to seven tasks, its interpretation reflects three reasoning tasks and four memory tasks rather than the distribution of requirements in a clinical application. This

analysis therefore separates configuration comparisons from assumptions about application value. Mitchell et al., 2019, advocate reporting performance under clearly specified conditions and intended uses. The findings extend that reporting logic by showing why parameter count, tuning category and task coverage must remain visible together. The distribution of matched differences exposes an additional evaluation problem: average deterioration can coexist with improvement on more than half the tasks within a configuration pair. Table 7 illustrates this pattern for the 13 and 70 billion parameter Llama 2 pairs, whose positive median differences coexist with negative mean differences. The analytical insight is that the frequency of improvement and the magnitude of deterioration answer different questions. Reporting either alone would obscure the performance profile. Evaluators should therefore present task specific differences, their direction counts and their magnitudes together, with particular attention to large declines on capabilities required by the intended application. This recommendation follows directly from the observed asymmetry rather than from an assumed clinical risk threshold. The transparency priorities articulated by van Dis et al., 2023, become operational here through inspectable comparisons that preserve contradictory results instead of compressing them into a single favourable or unfavourable assessment.

The international relevance lies in a transferable evaluation approach rather than a geographical performance claim. Tables 5 through 7 shows why institutions comparing available models should evaluate the configuration intended for use against the tasks it must perform, rather than select on family reputation, parameter count or conversational adaptation alone. Unlike the clinical workflow evaluation undertaken by Hager et al., the present evidence concerns published benchmark tasks; unlike the broader evaluation critique developed by Agrawal et al., it supplies a transparent configuration and task comparison within one fixed matrix. These contributions are complementary rather than interchangeable. Equation 1 offers a reproducible way to report matched accuracy differences, while the size comparisons identify where an apparent scaling advantage changes across tuning categories. Together, these findings refine capability and task alignment into an empirical reporting principle: retain the conditions under which an advantage appears, and do not generalize it across capabilities that produce different observed rankings.

7. Conclusion and Implications:

A medical language model's apparent advantage can disappear when the task changes. This analysis shows that development choices, computational capacity and conversational adaptation accompany conditional performance patterns rather than a stable hierarchy of capability. The core contribution is a comparative framework that preserves these conditions and exposes differences concealed by aggregate rankings. Improvements across several tasks can coexist with larger declines elsewhere, making the frequency and magnitude of performance changes distinct considerations. Theoretically, this evidence refines capability and task alignment by showing that configuration advantages require evaluation against specific demands. It establishes descriptive boundaries for broader explanations rather than identifying causal pathways. For decision makers, selection should begin with the capabilities required by the intended application, followed by direct comparison of the configurations under consideration. Procurement and governance processes should require task specific evidence, disclose substantial performance declines and avoid treating conversational adaptation or greater computational capacity as sufficient assurance of reliability. Operationally, evaluation routines should retain matched comparisons and report both typical differences and extreme deterioration. These principles offer institutions internationally a transparent basis for allocating evaluation resources and scrutinizing model claims. Their social relevance lies in supporting accountable decisions about medical information technologies without equating benchmark success with demonstrated patient benefit.

The available evidence offers specific opportunities for extension. Published aggregate results permit transparent configuration comparisons, while response level data would enable uncertainty estimation and examination of individual error patterns. Additional benchmark cohorts could establish whether the observed rankings persist across model generations. Controlled adaptation experiments could distinguish training effects from other configuration differences. Evaluations across languages, institutions and clinical workflows could assess transferability beyond the benchmark setting. Recording computational requirements alongside task performance would support comparisons of capability and resource demands. These extensions would test the stability and practical relevance of the conditional patterns identified here.

References:

1. Agrawal, M., Chen, I. Y., Gulamali, F., & Joshi, S. (2025). The evaluation illusion of large language models in medicine. npj Digital Medicine, 8, Article 600. https://doi.org/10.1038/s41746-025-01963-x
2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922
3. Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625-630. https://doi.org/10.1038/s41586-024-07421-0
4. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3458723
5. Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine, 30(9), 2613-2622. https://doi.org/10.1038/s41591-024-03097-1
6. 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. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596
7. Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2023). Med-HALT: Medical domain hallucination test for large language models. In Proceedings of the 27th Conference on Computational Natural Language Learning (pp. 314-334). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.21
8. Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226-1227. https://doi.org/10.1126/science.1213847

9. Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. PLOS Computational Biology, 9(10), Article e1003285. https://doi.org/10.1371/journal.pcbi.1003285
10. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180. https://doi.org/10.1038/s41586-023-06291-2
11. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930-1940. https://doi.org/10.1038/s41591-023-02448-8
12. van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. Nature, 614(7947), 224-226. https://doi.org/10.1038/d41586-023-00288-7