



Balsam: Decision Language Models

Maturity Benchmark

Report date: 11 October 2026



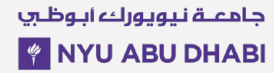
Benchmark Partners:



HUMAIN



Entropy



stc



groq

aiXplain

Message from the Secretary-General of the King Salman Global Academy for Arabic Language

The King Salman Global Academy for Arabic Language serves Arabic and strengthens its global presence through mutually reinforcing tracks: language planning, computing, education, and culture. Arabic's presence in AI technologies is now one facet of its standing among languages. The Academy has therefore focused on evaluating these technologies and their maturity in handling Arabic, believing that what gets measured gets improved.

The Academy launched the Balsam Benchmark with a first report evaluating generative language models through both human judgment and automatic evaluation, then expanded it in a second report with new questions and categories. This report takes that work a step further, extending the Benchmark to a class of language models not previously evaluated.

The Academy is pleased to publish the Decision Language Models Maturity Benchmark report, covering models that do not generate text but select a decision from specified options. The report asks not only whether the model's decision was correct, but also whether its confidence in that decision can be relied upon. It therefore uses four metrics: Accuracy, Brier Score (BS), Expected Calibration Error (ECE), and Area Under the Risk-Coverage Curve (AURC), across (37) tasks in (12) linguistic and cognitive categories. Its results are independently audited, strengthening their reliability for research and applied use.

As the circle of Benchmark partners widens with each report, the Academy invites academic and technical institutions, researchers, and interested parties to contribute to its next report by building its data, developing its methodology, and nominating models for evaluation. To collaborate with us, please contact the Balsam email address on the website.

In closing, we extend our thanks to His Highness Prince Badr bin Abdullah bin Farhan Al Saud — may God preserve him — Minister of Culture and Chairman of the Academy's Board of Trustees, for his continued support of the Academy's projects and work, and to the Benchmark partners who helped build its data and develop its methodology, strengthening the leading position of Arabic in the technical field.

Secretary-General
Prof. Dr. Abdullah bin Saleh Al-Washmi

Table of Contents

Decision Language Models Maturity Benchmark	4
General Evaluation Statistics for the Decision Language Models Maturity Benchmark	4
Task Categories Included in the Evaluation	5
Evaluation Methodology	8
Evaluation Method	8
Evaluation Metrics	9
Verification of the Results	9
Evaluated Models	10
Results	11
Overall Average of the Models	12
First: Accuracy	12
Second: Brier Score (BS)	13
Third: Expected Calibration Error (ECE)	13
Fourth: Area Under the Risk-Coverage Curve (AURC)	14
Model Size and Its Relation to Performance	14
Results by Categories and Tasks	15
Appendix (A): Definitions of Categories and Tasks	38

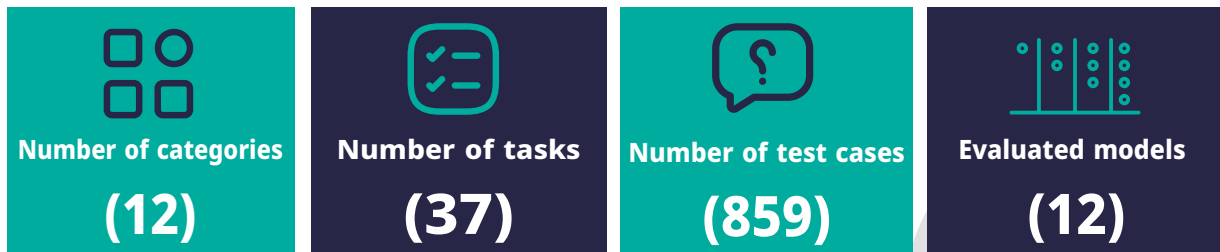
Decision Language Models Maturity Benchmark

The Decision Language Models Maturity Benchmark measures the ability of models to select a decision from among specified options in tasks written in Arabic, through multiple-choice questions covering a variety of linguistic and cognitive domains.

This report aims to present the complete official evaluation results as stored after verification. Four core metrics are displayed after being normalized from (0) to (100). An overall average with equal weights is derived from them, to provide a unified summary of the display scores, and this average is not regarded as an independent scientific metric.

Each test case consists of an input and context, depending on the task, instructions where available, a set of specified options, and one reference answer.

General Statistics on the Evaluation under the Decision Language Models Maturity Benchmark in the Current Release:



✓ Task Categories Included in the Evaluation:

Category	Category Name in Arabic	Task	Task Name in Arabic
Creative Writing	الكتابة الإبداعية	Text Completion	إكمال النصوص
		Text Continuation Evaluation	تقييم صحة تتمة النص
Entailment	التضمن	Semantic Similarity	التشابه الدلالي
		Textual Entailment	الاستلزام النصي
Factuality	دقة المعلومات	Answer Verification	التحقق من صحة الإجابة
		Claim Verification	التحقق من صحة الادعاء
Fill in the Blank	ملء الفراغ	Fill in The Blank	ملء الفراغ
Information Extraction	استخلاص المعلومات	Entity Categorization	تصنيف الكيانات
		Entity Relation Classification	تصنيف العلاقات بين الكيانات
		Extracting Required Information	استخلاص المعلومات المطلوبة

Category	Category Name in Arabic	Task	Task Name in Arabic
Logic	المنطق	Cause Effect Classification	تصنيف السبب والنتيجة
		Coherence Classification	تصنيف التماسك النصي
		Commonsense Validation	التحقق من المنطق السليم
		Logical Reasoning	الاستدلال المنطقي
		Sentence Ordering	ترتيب الجمل
Program Execution	تنفيذ البرنامج	Program Execution	تنفيذ البرامج
Question Answering	الإجابة عن الأسئلة	Answering Given Question	الإجابة عن الأسئلة المطروحة
Reading Comprehension	الفهم القرائي	Answerability Classification	تصنيف إمكان الإجابة
		Dialogue	اختيار الحوار المناسب
		Explanation	اختيار التفسير المناسب
		Question	فهم السؤال
		Textual Inference Validation	التحقق من الاستدلال النصي
		Part-of-Speech Tagging	تحديد القسم الكلامي

Category	Category Name in Arabic	Task	Task Name in Arabic
Text Classification	تصنيف النصوص	Dialect	تحديد اللهجة
		Dialogue Act Recognition	تعرف أفعال الحوار
		Emotion Detection	اكتشاف المشاعر
		Ethics Classification	تصنيف السلوكيات الأخلاقية
		Offensive Language Detection	اكتشاف الإساءة
		Question Categorization	تصنيف الأسئلة
		Sarcasm Detection	اكتشاف السخرية
		Sentiment Analysis	تحليل المشاعر
		Spam Detection	كشف الرسائل المزجة
		Stereotype Detection	اكتشاف القوالب النمطية
		Text Categorization	التصنيف الفئوي للنصوص
		Topic Identification	تحديد الموضوع

Category	Category Name in Arabic	Task	Task Name in Arabic
Text Manipulation	التعديل على النصوص	Grammar Correction	تصحيح القواعد النحوية
Translation/Transliteration	الترجمة الآلية والنقل الصوتي	Machine Translation	الترجمة الآلية

Evaluation Methodology:

A unified evaluation methodology was adopted to measure the maturity of decision-making language models across a suite of Arabic tasks. Each model is prompted to select one option from a specified set of choices for every test case. Each case includes a context or input, instructions tailored to the task's nature, a candidate set of options, and a single reference answer. A model's prediction is deemed correct when it selects the option matching the reference answer.

The official evaluation is based on test data from the second Balsam release (Balsam V2), excluding development data from the final results. The current release comprises 859 test cases distributed across 37 tasks within 12 categories, with all test cases in this release being eligible for evaluation.

Evaluation Method:

The evaluation process relies on a dedicated adapter for each model. This adapter converts model outputs into a standardized decision format prior to evaluation, allowing all models to be evaluated consistently regardless of their execution mechanism. The output includes the option selected by the model alongside the probability distribution over the available options when natively provided by the model. Probability distributions are not derived from a single confidence score if not natively supported.

Results are calculated at three levels: the overall benchmark score, category-level results, and task-level results. The overall score and category results are calculated directly at the test-case level rather than taking a simple average of task results, thereby preserving the proportional weight of test cases within each group.

Evaluation Metrics:

The methodology relies on four core metrics to evaluate model performance and outputs. To make the results easier to read and compare across models, all scientific values are converted into a unified display score ranging from 0 to 100, where a higher score always indicates better performance and higher quality.

Accuracy

Measures the proportion of correct decisions or answers made by the model out of all test cases. Its display score is calculated by multiplying the scientific value of the metric by 100, according to the formula: $\text{Display score} = \text{scientific value} \times 100$

Brier Score (BS)

Measures the quality of the probability distribution the model assigns to the options and how close it is to the correct answer, not only whether the final choice is correct. Because the scientific value of this metric expresses an error rate, where a lower value is better, it is converted to the display score according to the formula: $\text{Display score} = (1 - \text{scientific value}) \times 100$

Expected Calibration Error (ECE)

Measures how realistic and reliable the confidence levels the model assigns to its decisions are compared with its actual accuracy, showing how well the model's confidence agrees with the correctness of its decisions. Because the scientific value of this metric expresses the amount of calibration error, where a lower value is better, it is converted to the unified display score according to the formula: $\text{Display score} = (1 - \text{scientific value}) \times 100$

Area Under the Risk-Coverage Curve (AURC)

Measures the ability of the model's confidence scores to rank decisions by their reliability, so that the more reliable and correct decisions come before the others. The display score is calculated from the scientific value of the area under the risk-coverage curve, where a lower scientific value is better, according to the formula: $\text{Display score} = (1 - \text{scientific value}) \times 100$

Verification of the Results:

The official results are automatically verified before approval: the test cases are complete, with none missing or duplicated; the data release matches its fingerprint; and the probability distributions are valid where available. The metrics are then recomputed from the case records and checked against the stored values.

The four core metrics are shown as display scores from 0 to 100, where higher is always better. This scaling does not change how the scientific metrics are calculated; it only unifies their direction for comparison. The overall average is the equal-weighted mean of the four display scores (Accuracy, Brier Score, ECE, and AURC). It is a summary for display, not an independent scientific metric, and it does not replace the original values.

Evaluated Models:

Model Name	Source Availability	Size (Number of Parameters)	Description	Reference / Link
Bespoke Labs — Bespoke-Nimble-9B	Open source	9,000,000,000	A model specialized in evaluating context-related options and deciding among them without generating a free-form answer.	Link
Cloudflare — Clef	Open source + official API	27,000,000,000	A decision model hosted through Cloudflare Workers AI. It chooses among specified options and returns their probabilities.	Link
Convai Innovations — Laya Multilingual	Open source	322,000,000	A multilingual, non-generative decision model designed to choose among specified options and return their probabilities. It supports Arabic among the languages it covers.	Link
Decider-2B	Open source	1,881,825,088	A model specialized in deciding among specified options without generating text.	Link
Faisal	Open source	4,000,000,000	An Arabic decision model that relies on calibrated confidence in its outputs.	Link
IJyad — Jeb	Open source	177,609,985	A non-generative Arabic decision model designed for classification, inference, and decision-making tasks.	Link
Jared Palmer — Kev-0.8B	Open source	800,000,000	A small, non-generative decision model designed to choose among specified options without generating text.	Link
TypeSafe Jev — jev-1.13.0	Closed source	Not available	A model specialized in making structured decisions. It supports selection, scored evaluation, and binary judgments without generating text.	Link
Wouze — Laya-Ara	Weights available for research / non-commercial use	321,908,998	A non-generative Arabic decision model designed for Arabic language understanding, classification, and inference.	Link

Liquid AI — d1-3B	Open source	3,123,483,888	An open-weight local decision model from Liquid AI. It chooses among specified options and returns their probabilities in a single pass.	Link
Microsoft — Microsoft-Decision-1	Not publicly disclosed (available via API)	9,000,000,000	A non-generative decision model from Microsoft, post-trained on Qwen3.5-9B, that chooses among specified options and returns a calibrated probability for each option.	Link
OpenAI — GPT-6 Luna (Decisions)	Closed source	Not available	A decision model from OpenAI, evaluated via the Decisions API, that chooses among specified options and returns their probabilities.	Link

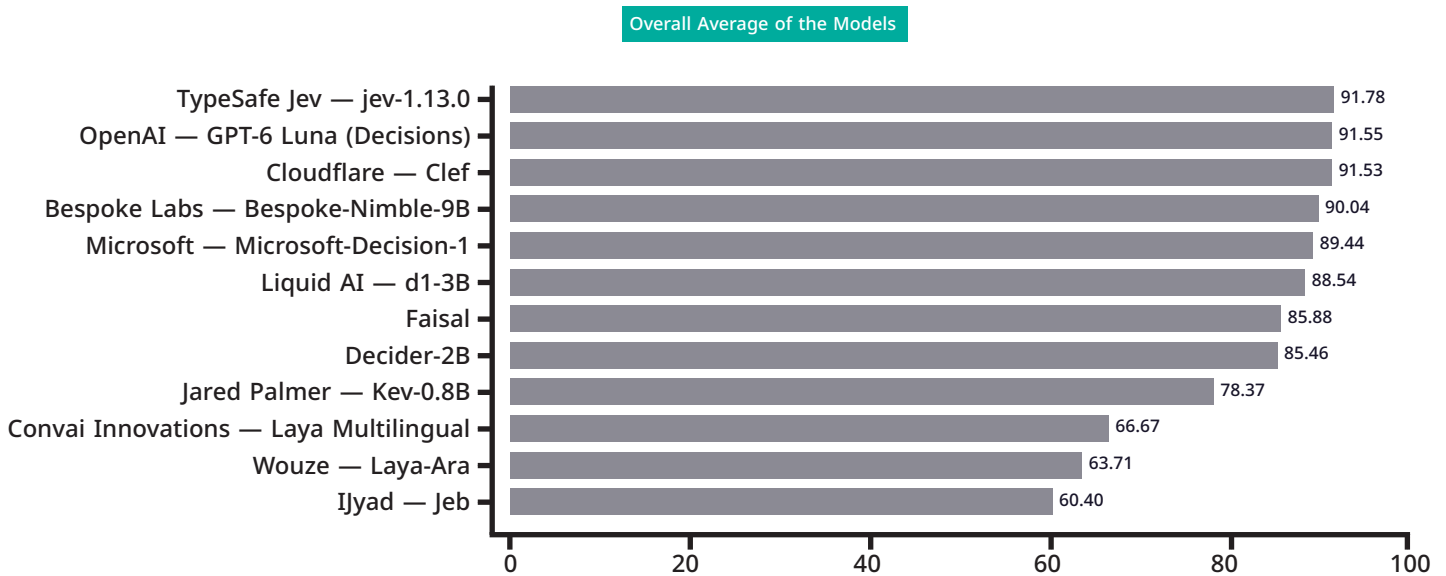
Note: Clef was evaluated through the Cloudflare Workers AI API, which provides a hosted version of the same open-weight Clef model. The results reported here reflect its execution via the API.

Results:

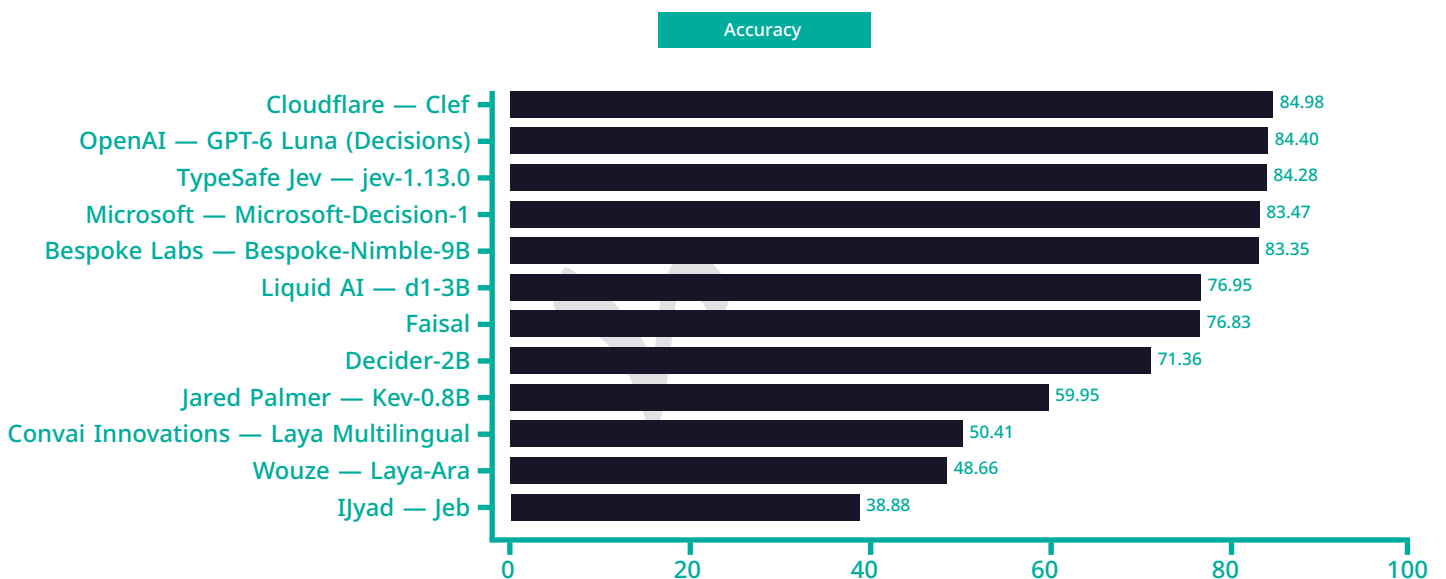
Model	Accuracy	Brier Score (BS)	Expected Calibration Error (ECE)	Area Under the Risk-Coverage Curve (AURC)	Overall Average
TypeSafe Jev — jev-1.13.0	84.28	91.08	96.36	95.40	91.78
OpenAI — GPT-6 Luna (Decisions)	84.40	91.58	95.73	94.47	91.55
Cloudflare — Clef	84.98	91.08	95.71	94.33	91.53
Bespoke Labs — Bespoke-Nimble-9B	83.35	90.08	93.26	93.48	90.04
Microsoft — Microsoft-Decision-1	83.47	89.99	91.49	92.80	89.44
Liquid AI — d1-3B	76.95	88.28	97.49	91.45	88.54
Faisal	76.83	86.76	89.12	90.82	85.88
Decider-2B	71.36	86.08	97.32	87.09	85.46
Jared Palmer — Kev-0.8B	59.95	81.91	97.23	74.39	78.37
Convai Innovations — Laya Multilingual	50.41	76.13	84.02	56.14	66.67
Wouze — Laya-Ara	48.66	75.06	79.24	51.89	63.71
IJyad — Jeb	38.88	75.21	84.75	42.77	60.40

All scores are from (0) to (100), and a higher value is better. The overall average is the equal-weighted arithmetic mean of the Accuracy, Brier Score (BS), Expected Calibration Error (ECE), and Area Under the Risk-Coverage Curve (AURC) scores.

Overall Average of the Models



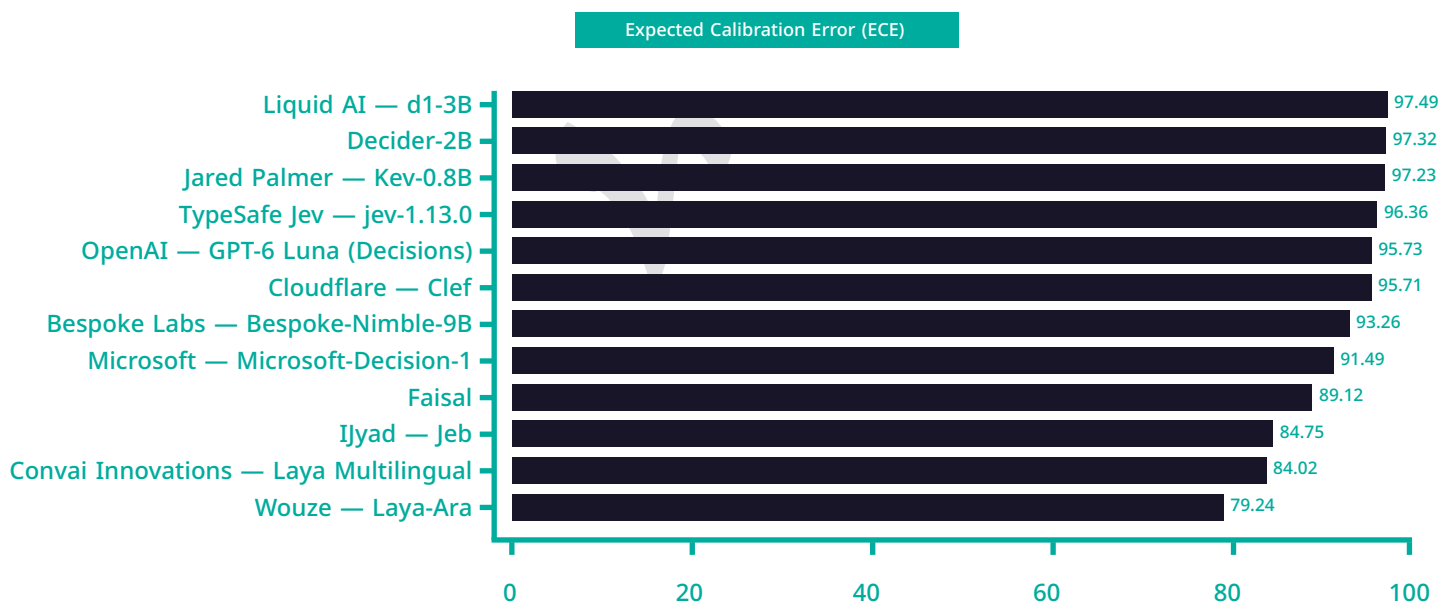
First: Accuracy



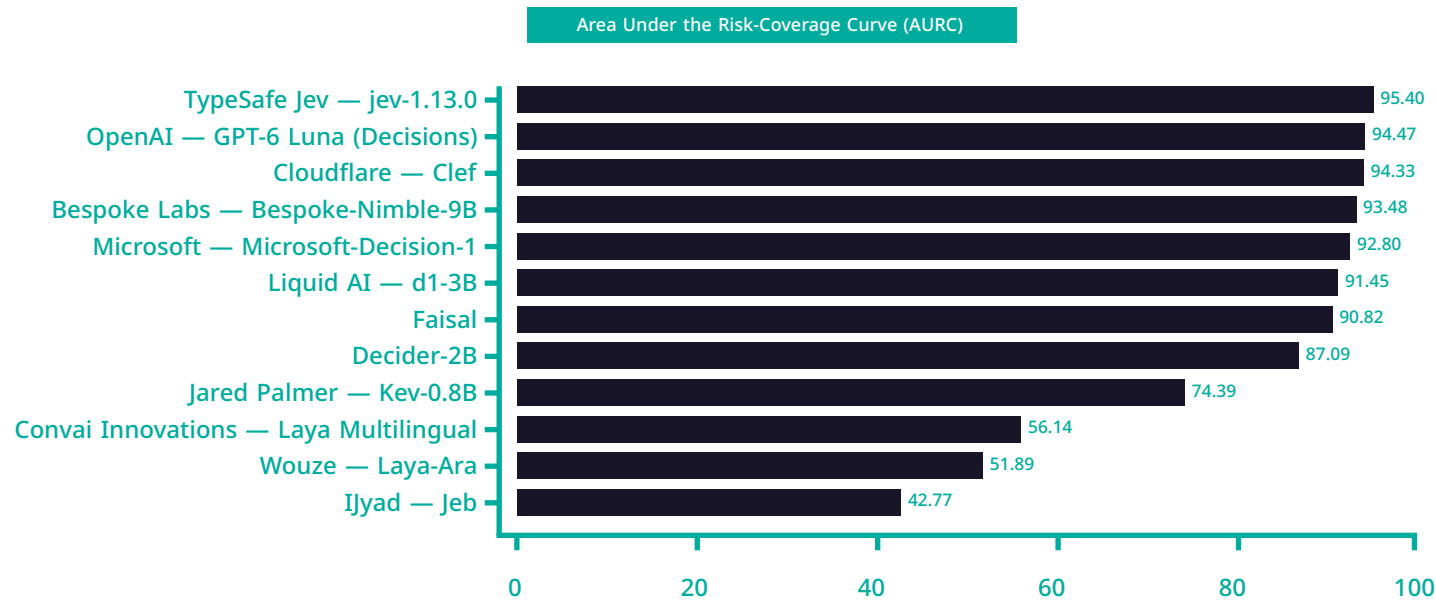
Second: Brier Score (BS)



Third: Expected Calibration Error (ECE)



Fourth: Area Under the Risk-Coverage Curve (AURC)

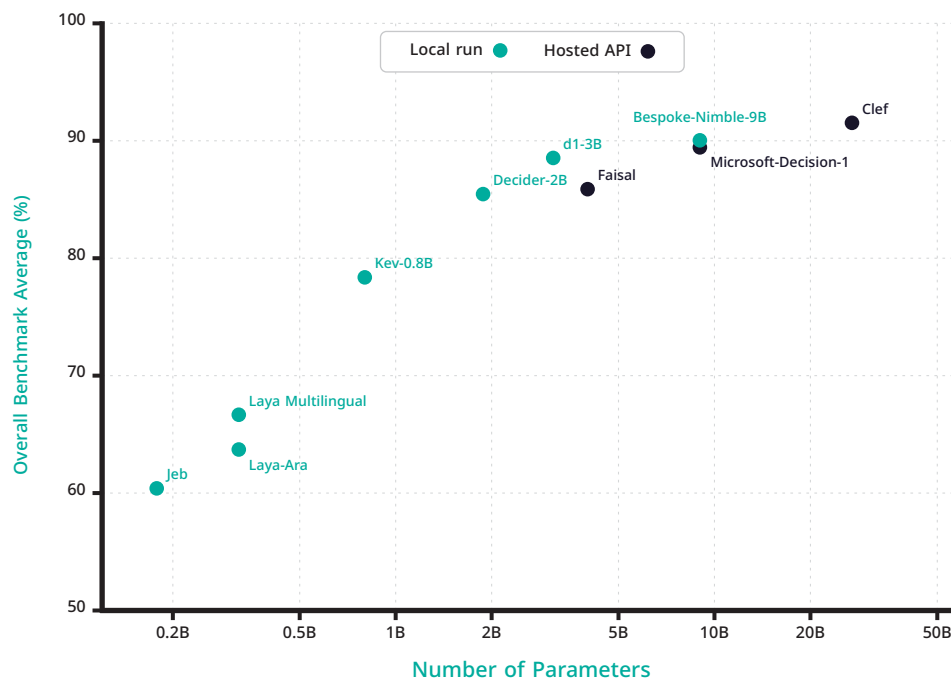


Model Size and Its Relation to Performance

The table shows the sizes of the models included in the Benchmark according to documented and available information, along with their evaluation method and overall average performance. Models with no documented size information are shown as “Undisclosed”.

Model	Model Size (Parameter Class)	Evaluation Method	Overall Average
TypeSafe Jev — jev-1.13.0	Undisclosed	Hosted API	91.78
GPT-6 Luna (Decisions)	Undisclosed	Hosted API	91.55
Cloudflare — Clef	27B	Hosted API	91.53
Bespoke-Nimble-9B	9B	Local run	90.04
Microsoft-Decision-1	9B	Hosted API	89.44
Liquid AI — d1-3B	3.12B	Local run	88.54
Faisal	4B	Hosted API	85.88
Decider-2B	1.88B	Local run	85.46
Kev-0.8B	0.8B	Local run	78.37
Laya Multilingual	0.322B	Local run	66.67
Laya-Ara	0.322B	Local run	63.71
Ijyad — Jeb	0.178B	Local run	60.40

Note: Size values represent the model size class or the documented number of parameters according to the information available for each model; for customized models such as Faisal, the value refers to the size of the base model. For Microsoft — Microsoft-Decision-1, the value (9B) corresponds to its base model (Qwen3.5-9B). The following chart is limited to models with documented size information.



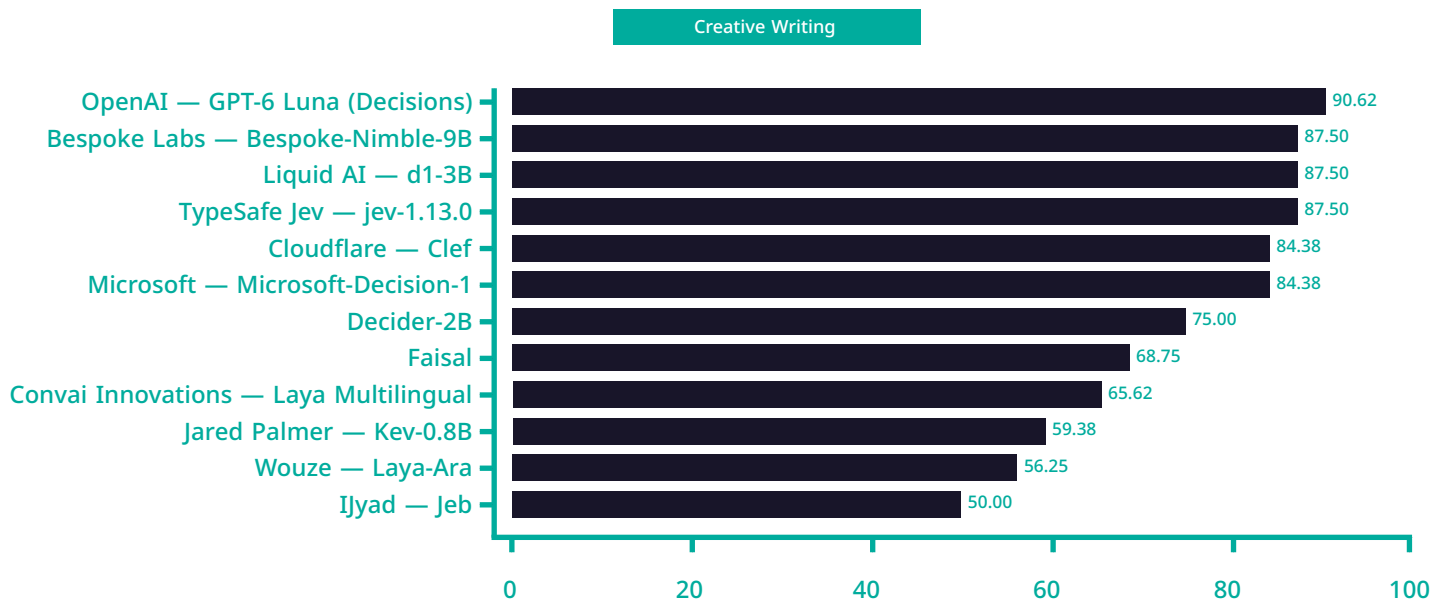
The figure shows the relationship between model size, represented by the number of parameters, and its overall average performance in the Benchmark. The results show a general trend of higher performance with larger model size; however, this trend does not apply consistently to all models, as some smaller models outperformed larger ones. This indicates that model size alone does not explain the level of performance, which may be influenced by other factors such as model architecture, training methodology, and task specialization. The figure also shows how the models were run during evaluation, either locally or through a Hosted API. The chart is limited to models with a disclosed number of parameters or a documented base model size.

Results by Category and Task:

The following task charts show the Accuracy display score from (0) to (100) for each task. Higher is better, and the higher value appears at the top of the chart. The category result, however, is taken directly from its stored summary over its items, not from an average of the task scores. For the results of the other metrics for each category, please visit [the report's interactive interface](#).

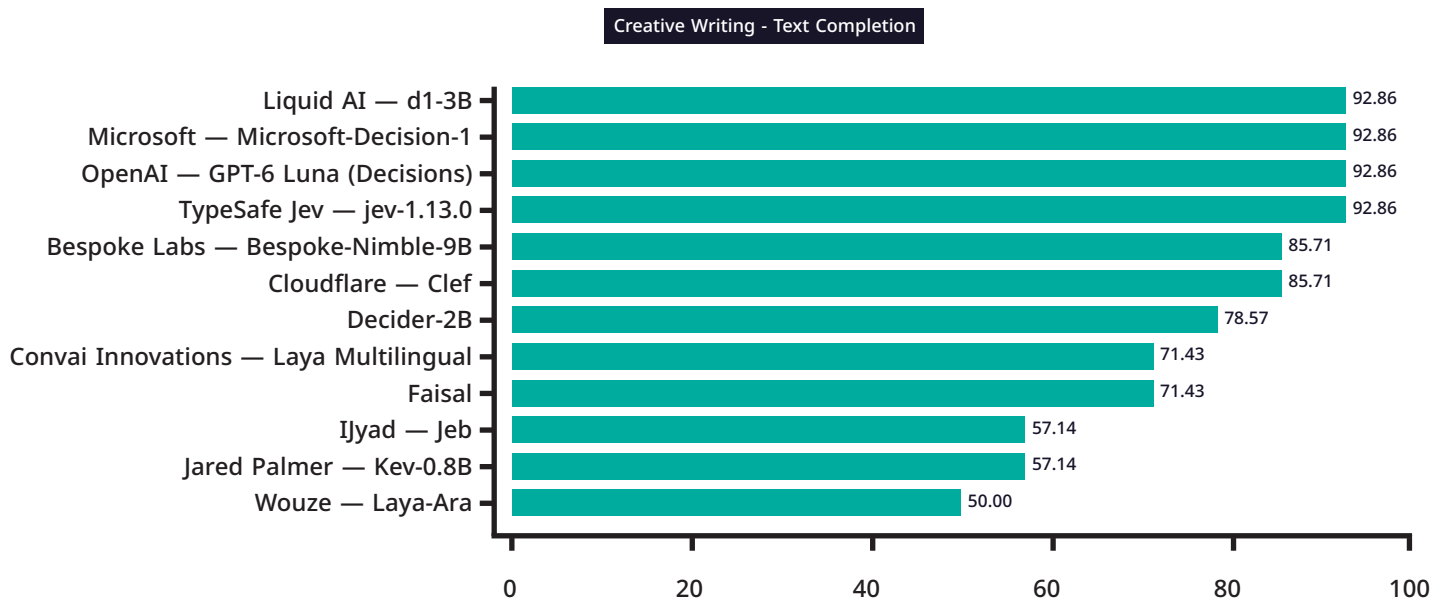


1. Creative Writing

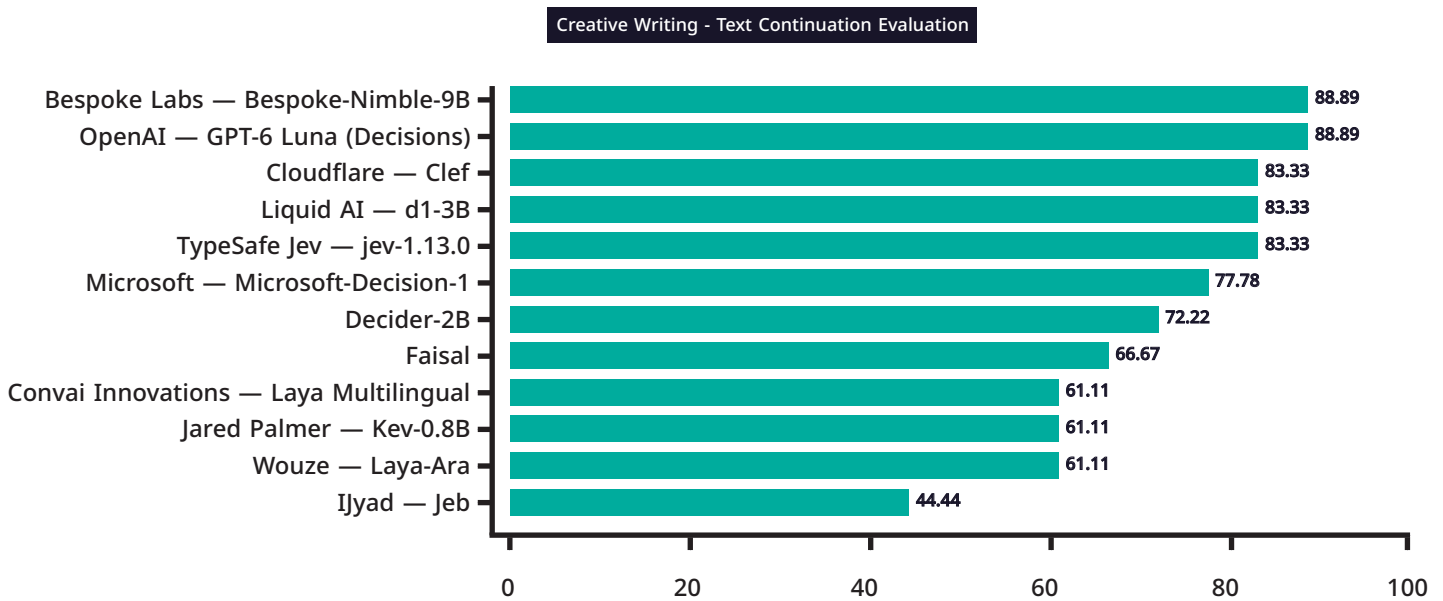


Subtasks:

1.1 Text Completion



1.2 Text Continuation Evaluation

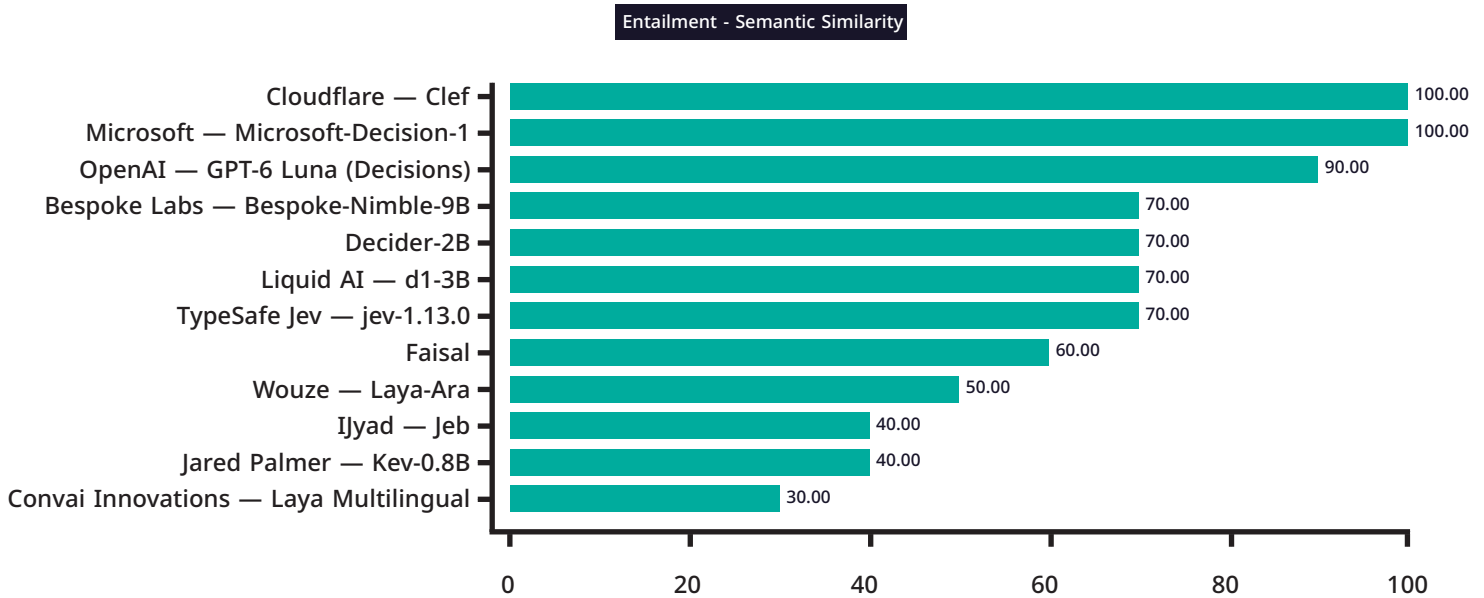


2. Entailment

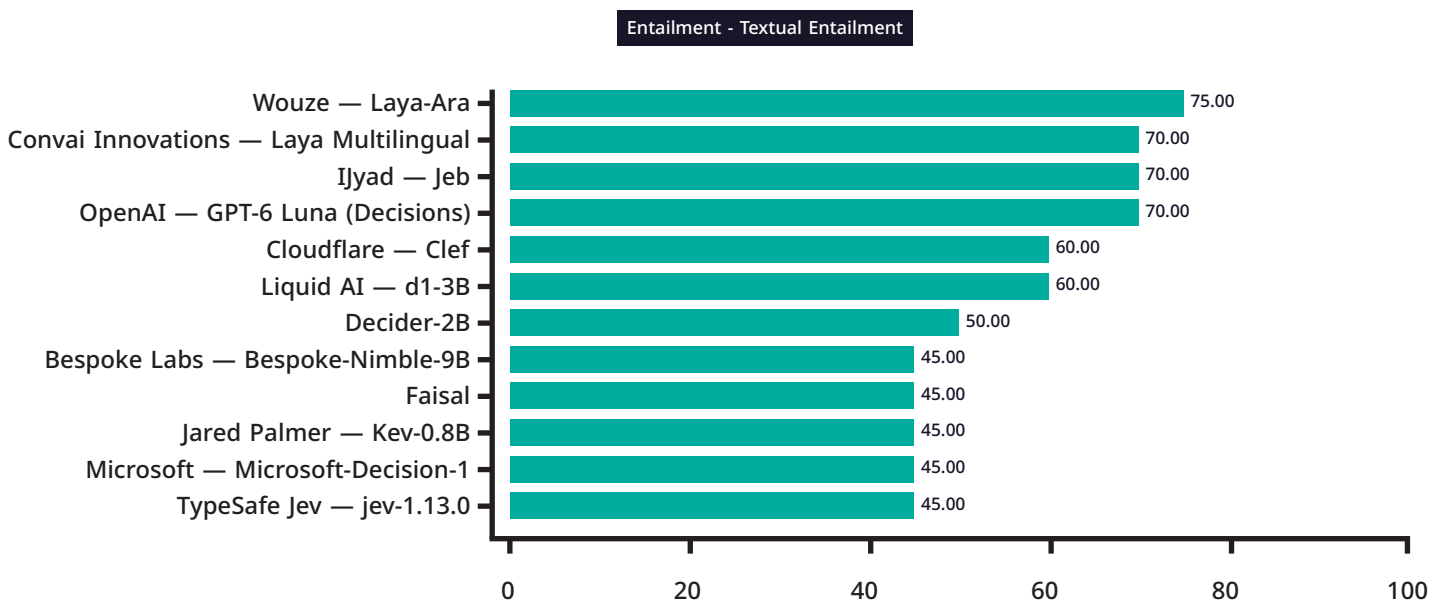


Subtasks:

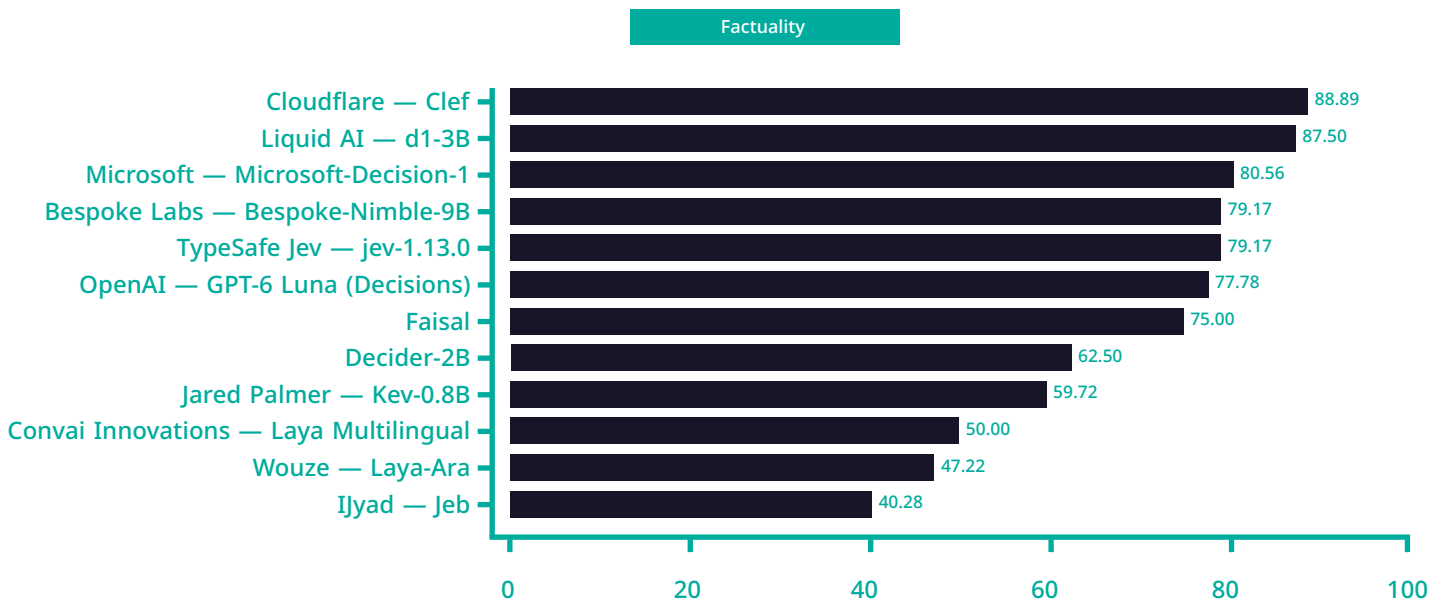
2.1 Semantic Similarity



2.2 Textual Entailment

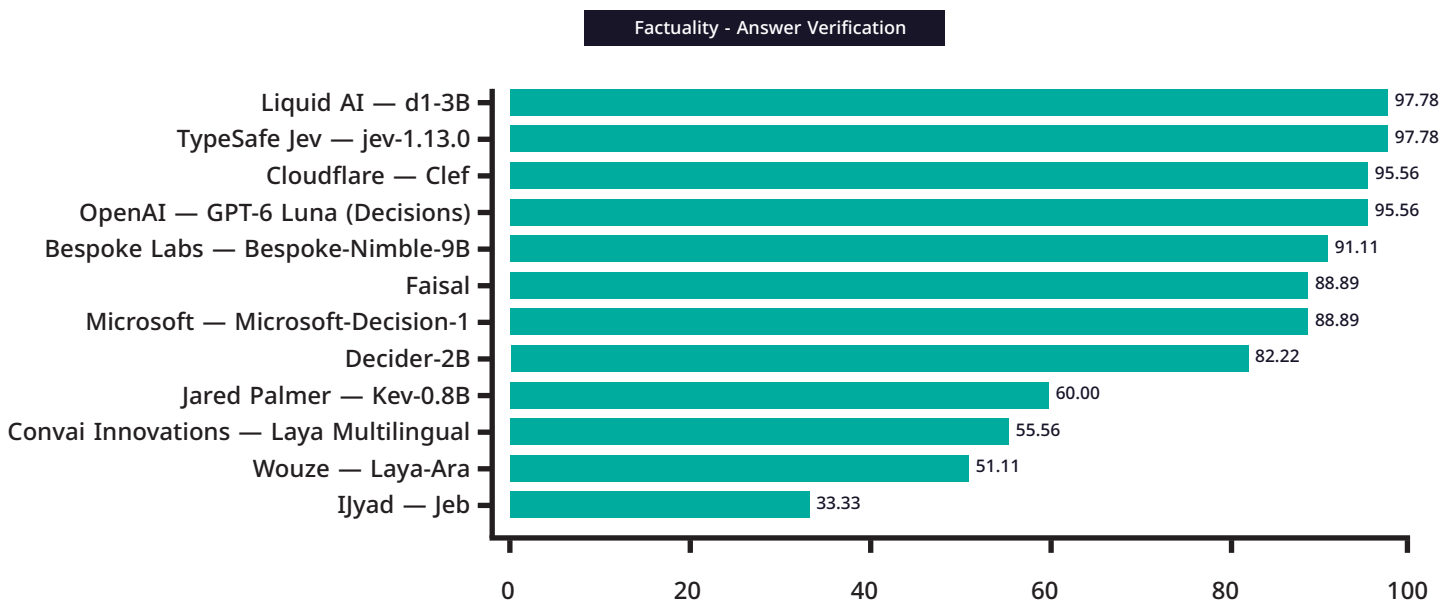


3. Factuality

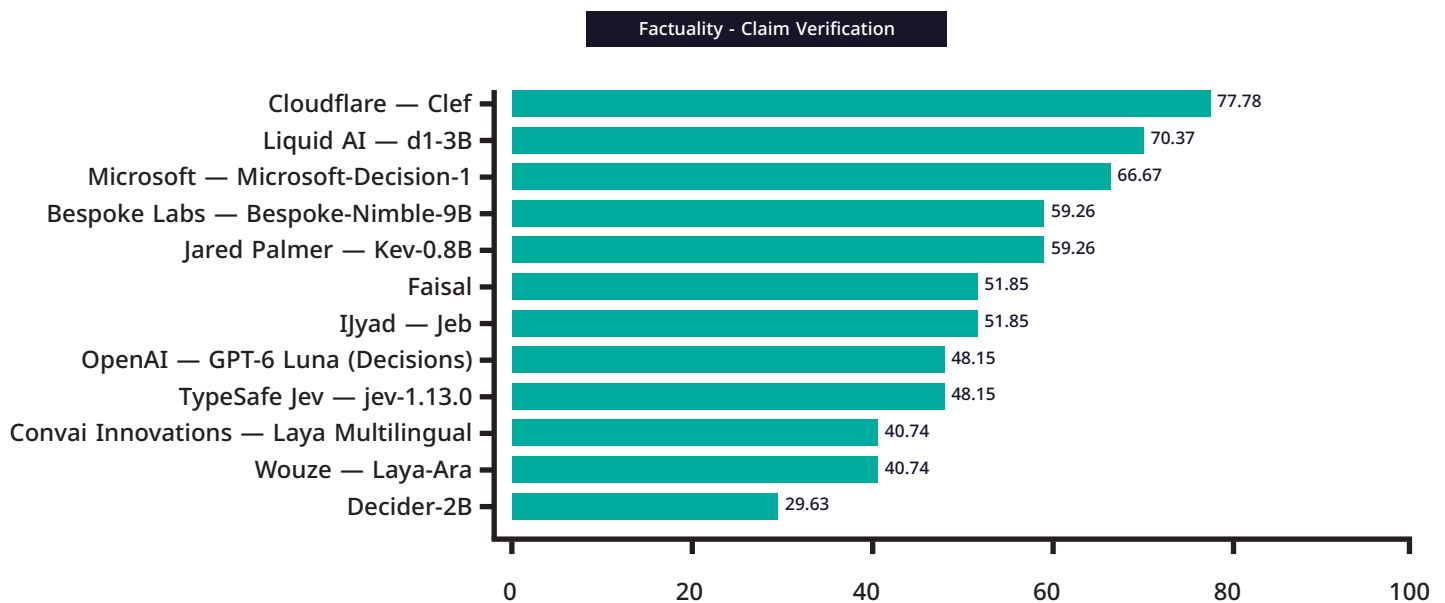


Subtasks:

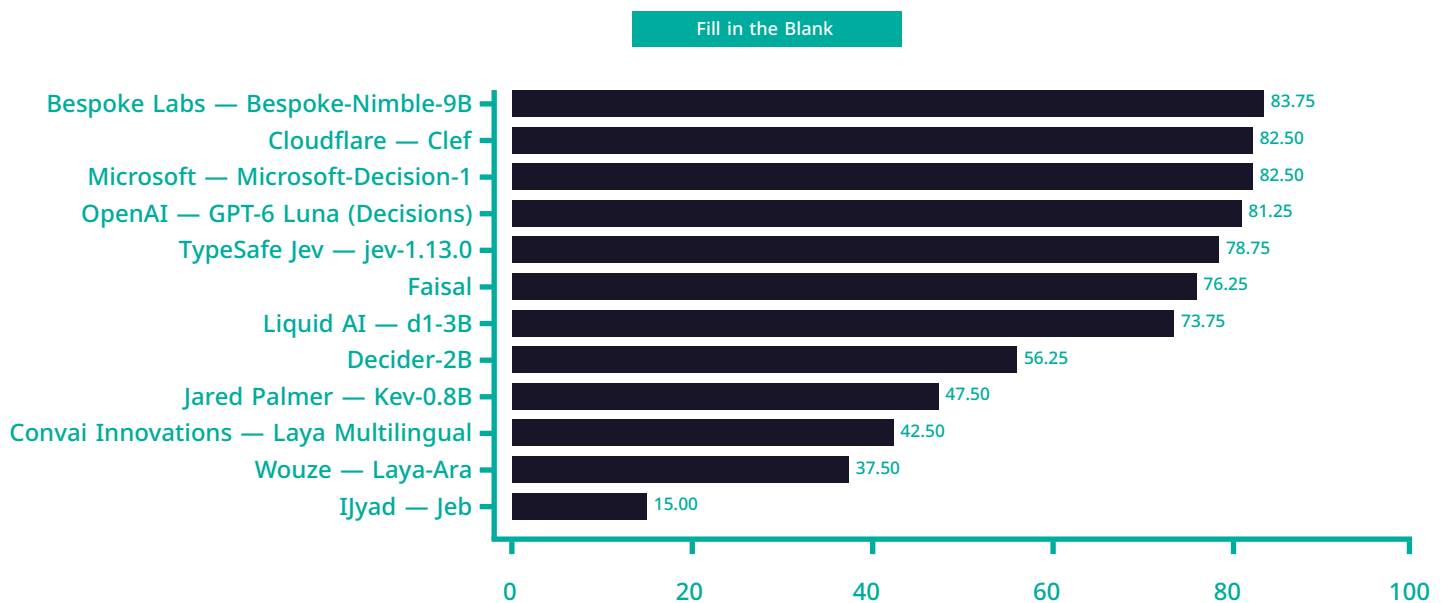
3.1 Answer Verification



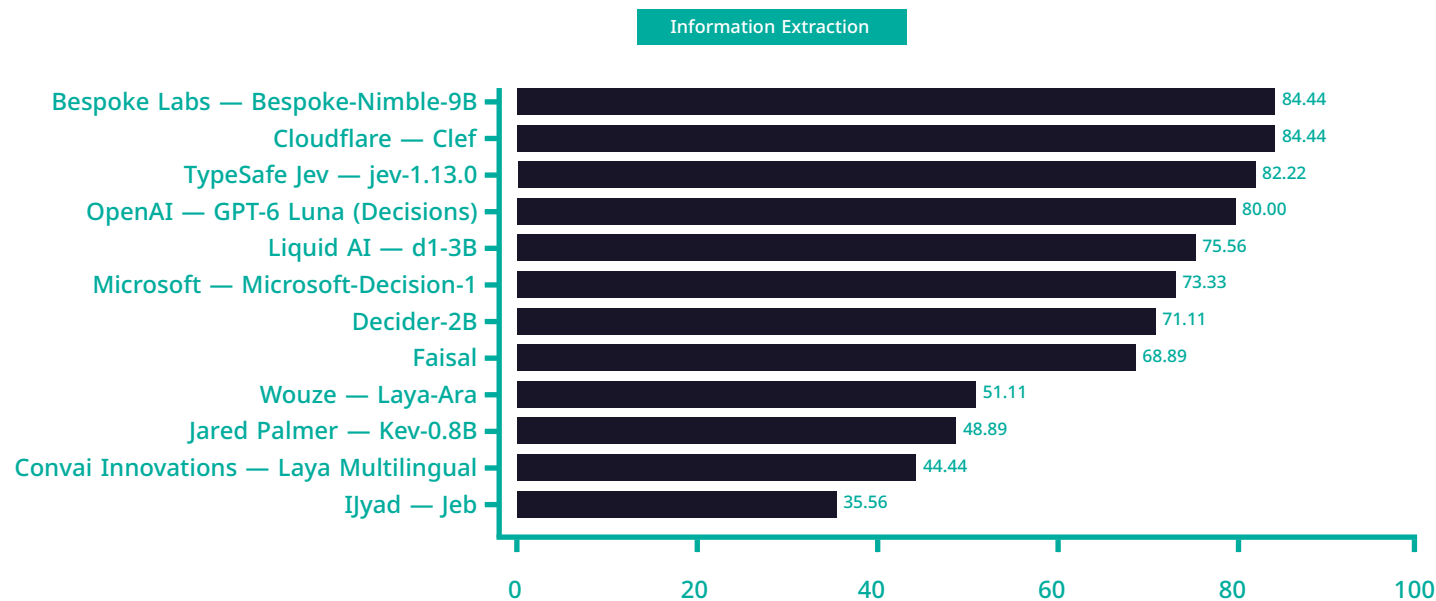
3.2 Claim Verification



4. Fill in the Blank

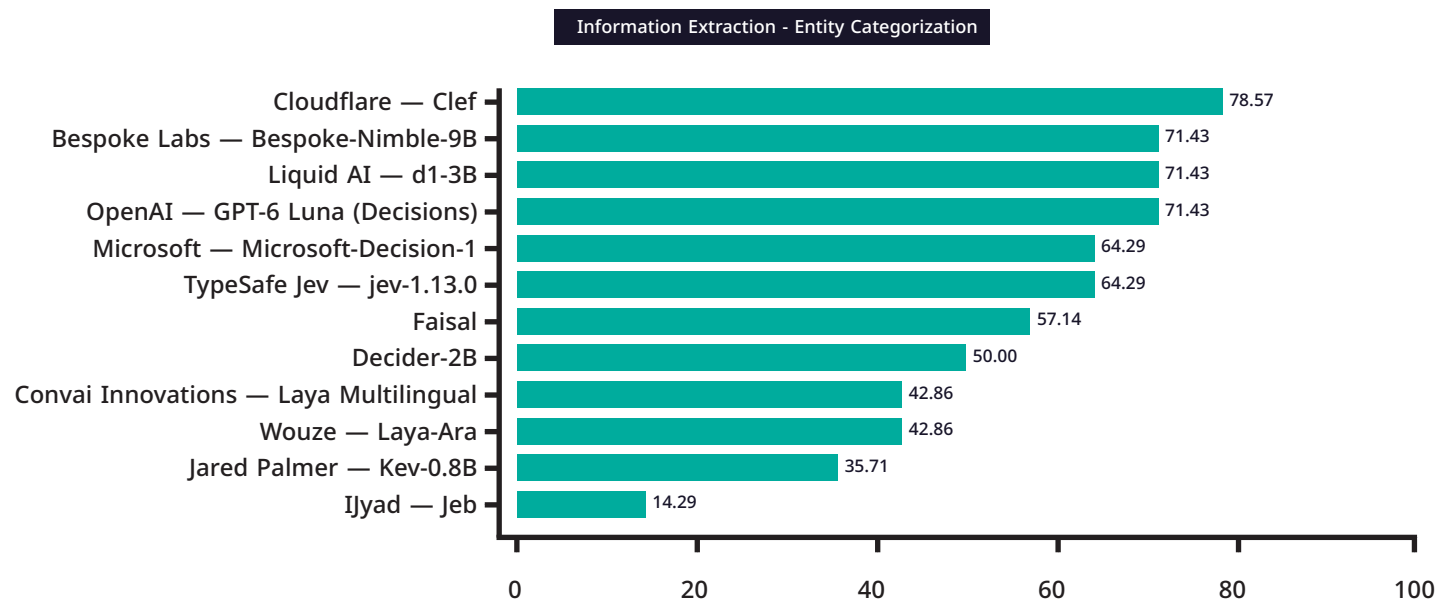


5. Information Extraction

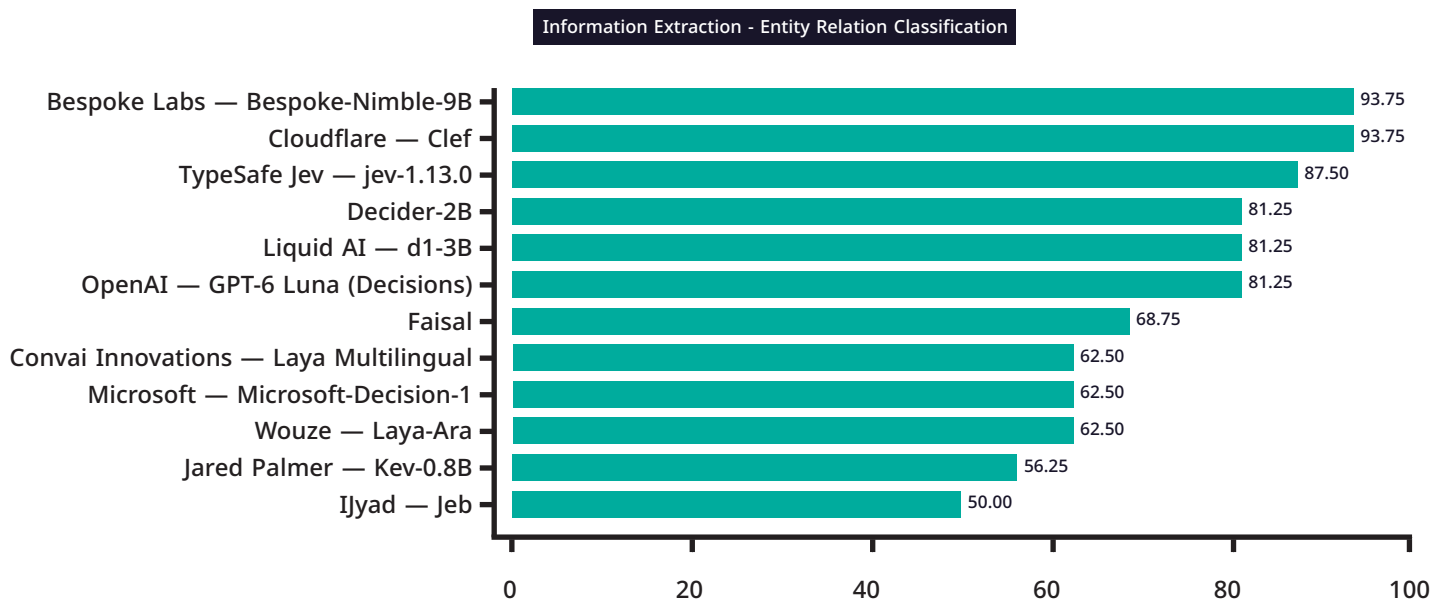


Subtasks:

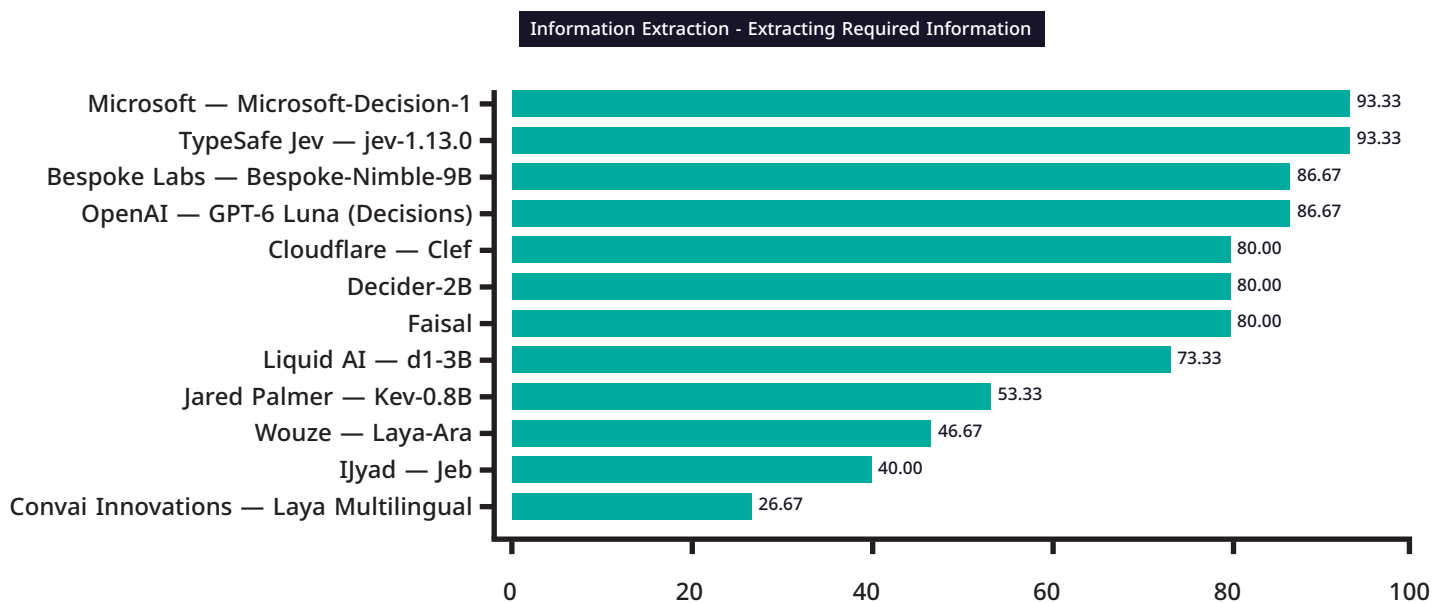
5.1 Entity Categorization



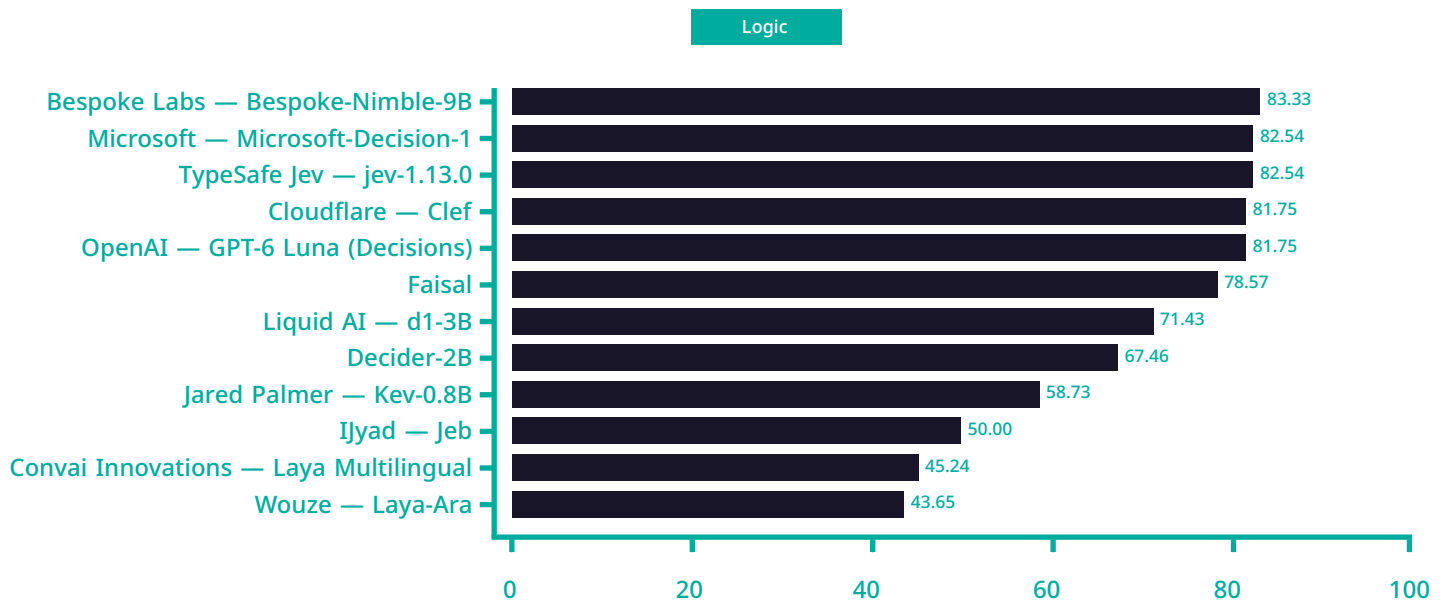
5.2 Entity Relation Classification



5.3 Extracting Required Information

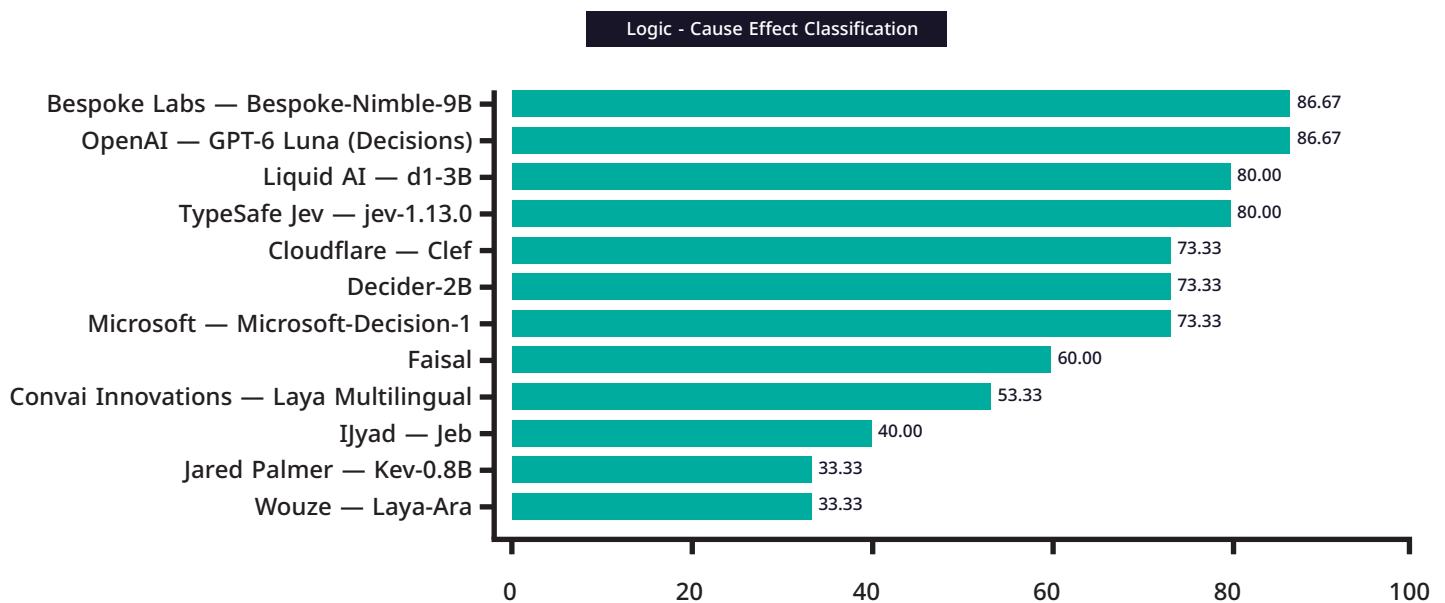


6. Logic

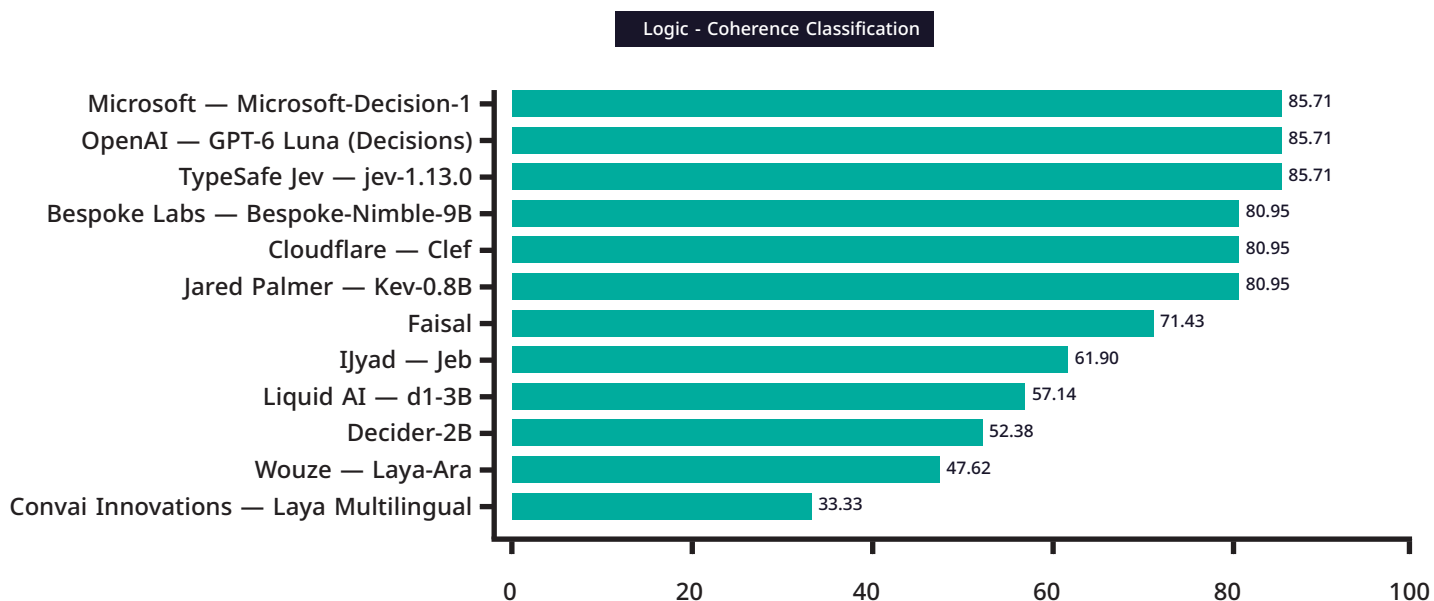


Subtasks:

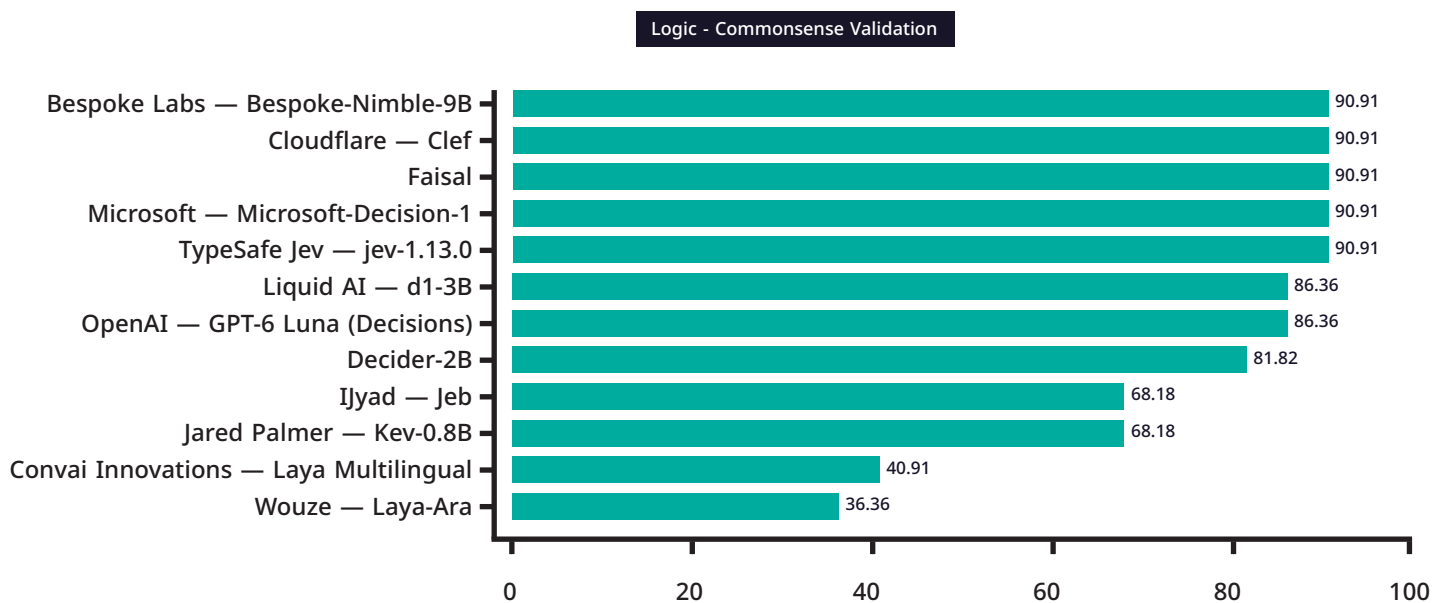
6.1 Cause Effect Classification



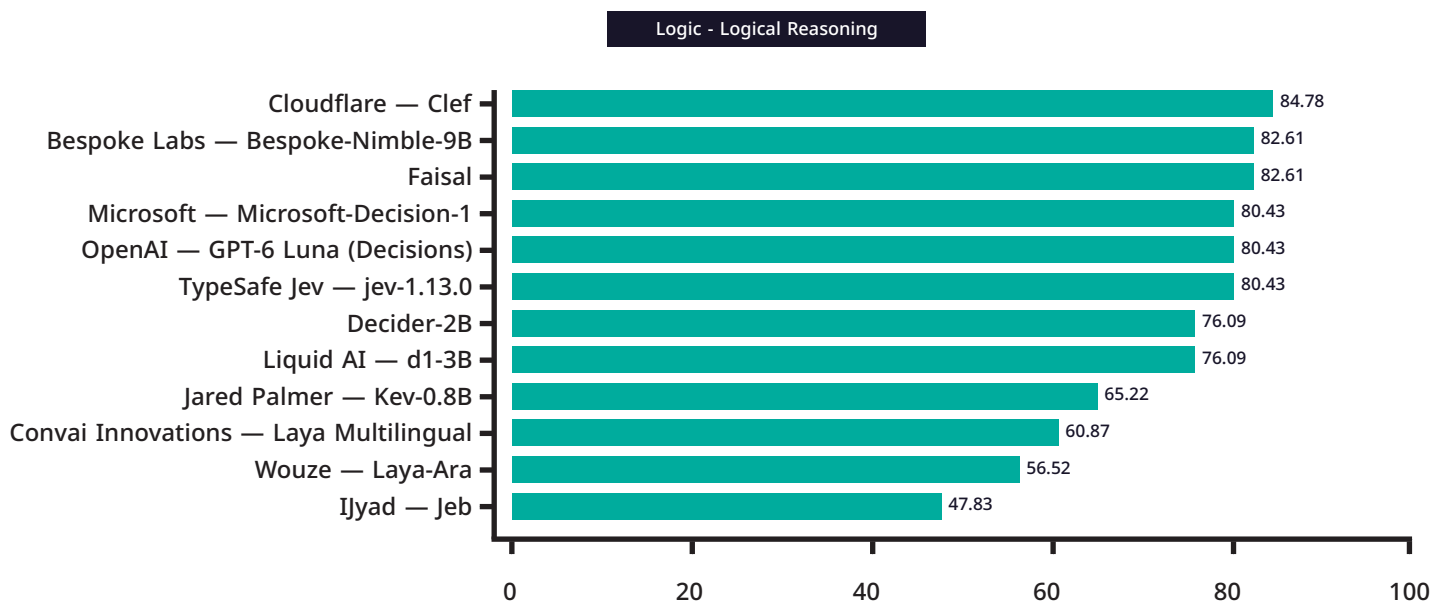
6.2 Coherence Classification



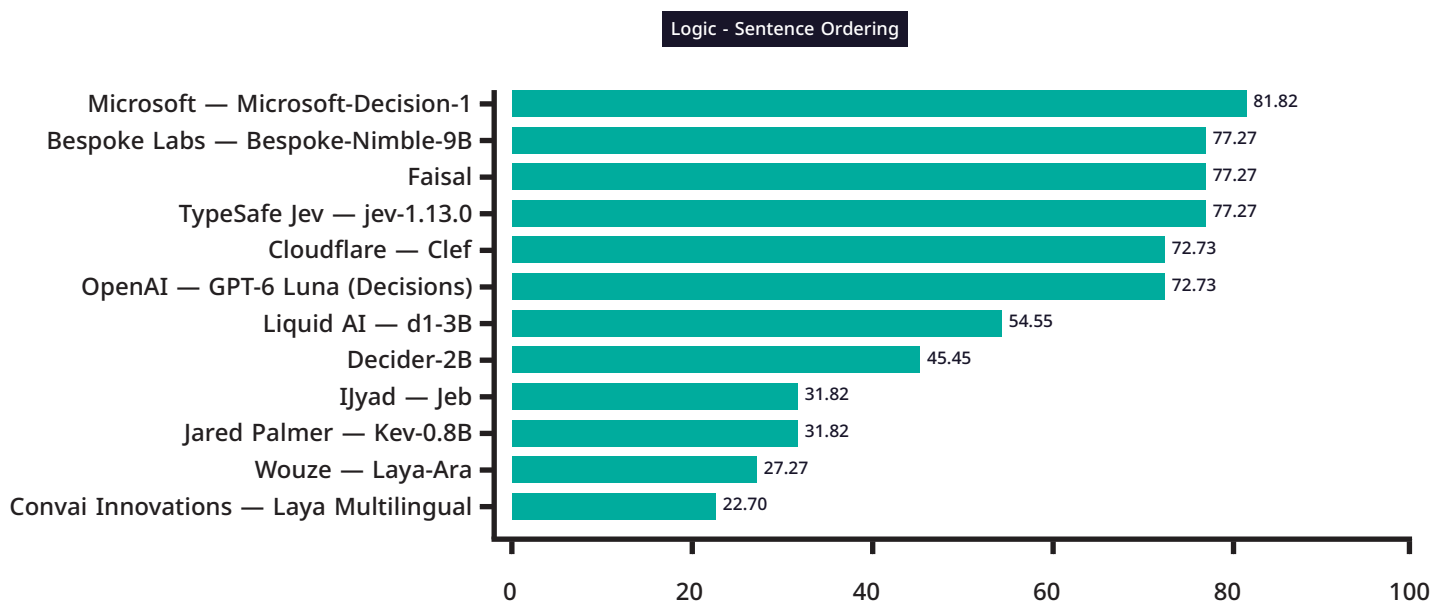
6.3 Commonsense Validation



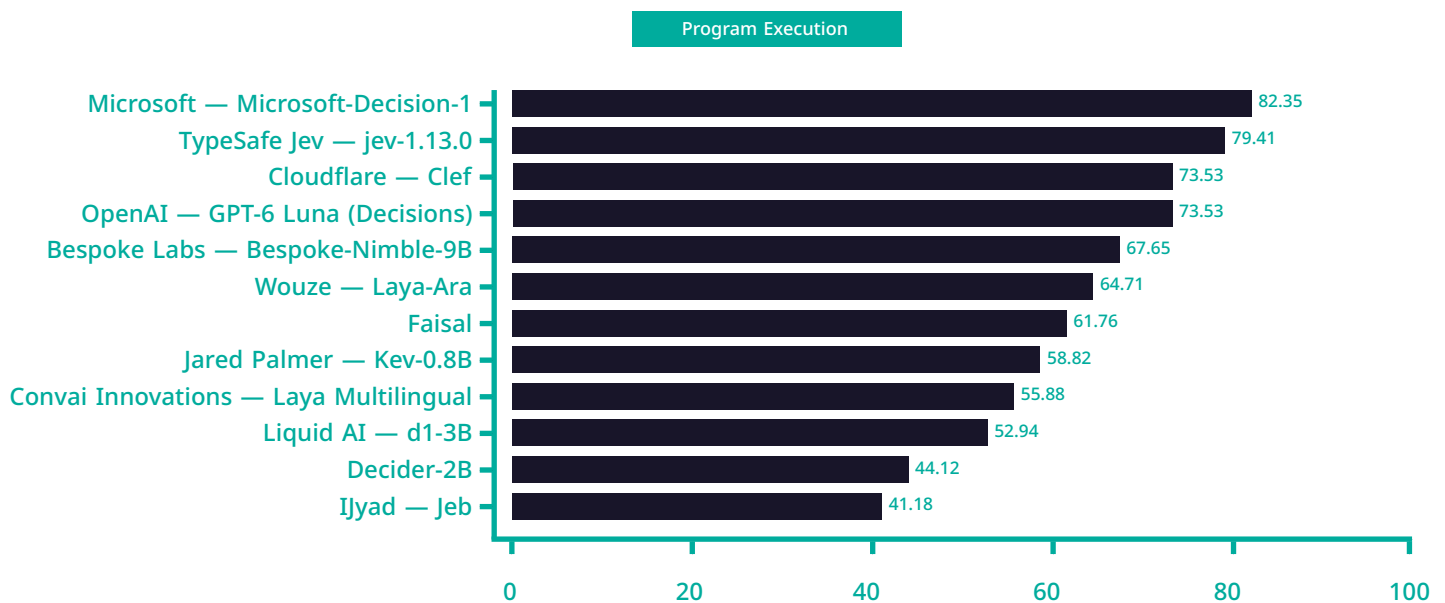
6.4 Logical Reasoning



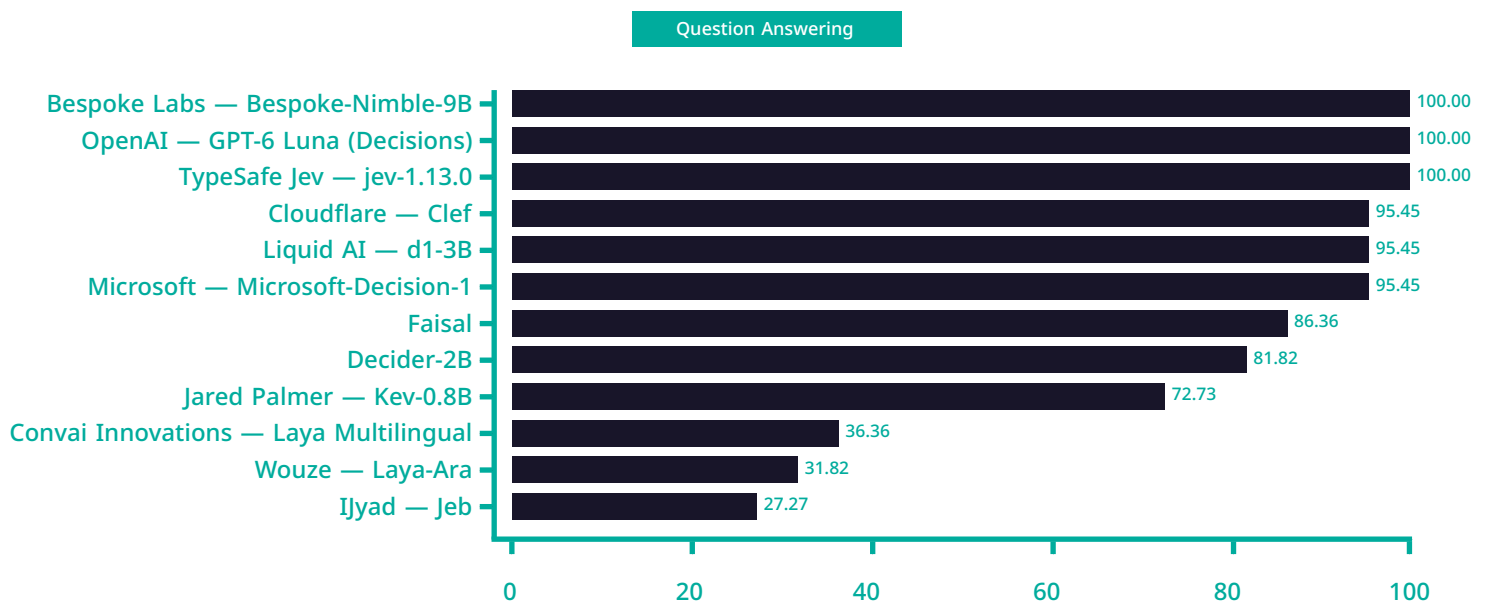
6.5 Sentence Ordering



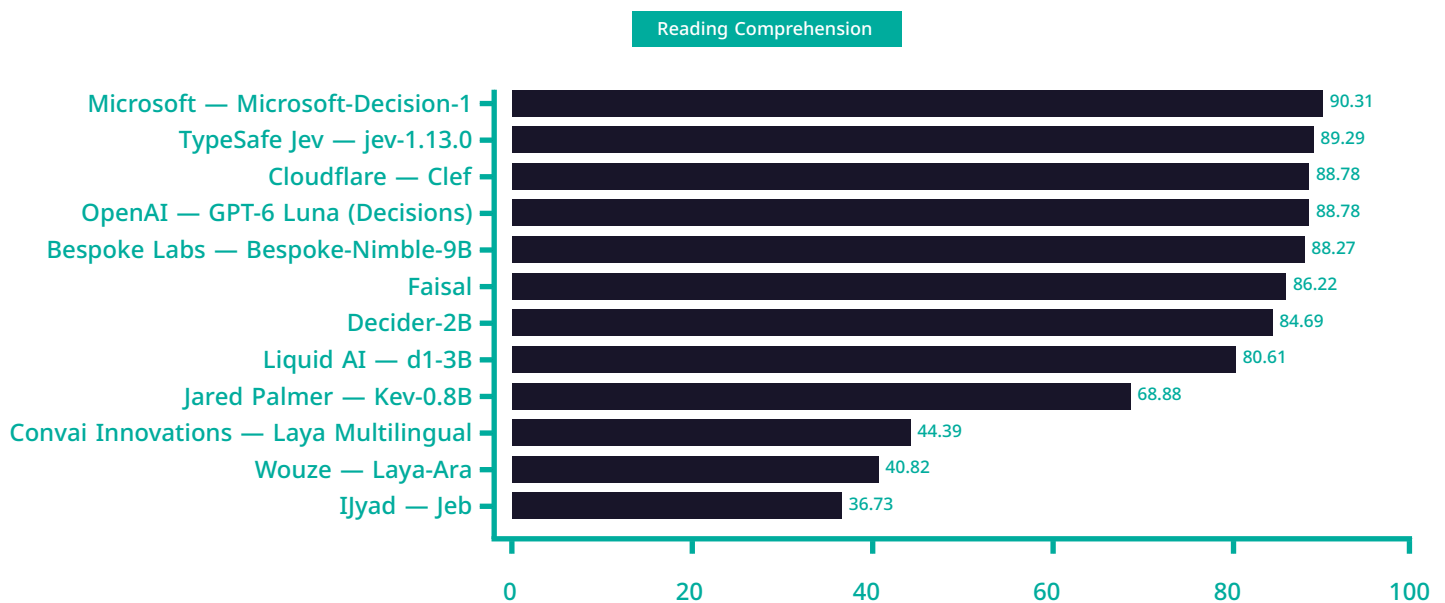
7. Program Execution



8. Question Answering

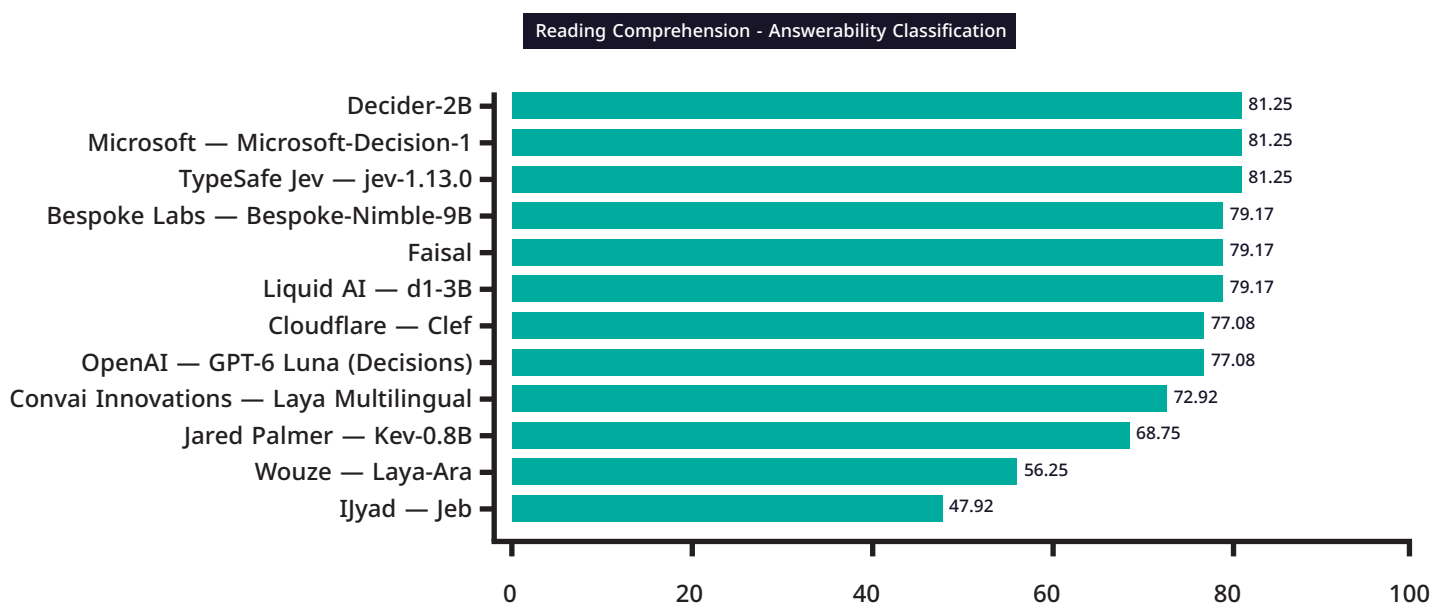


9. Reading Comprehension

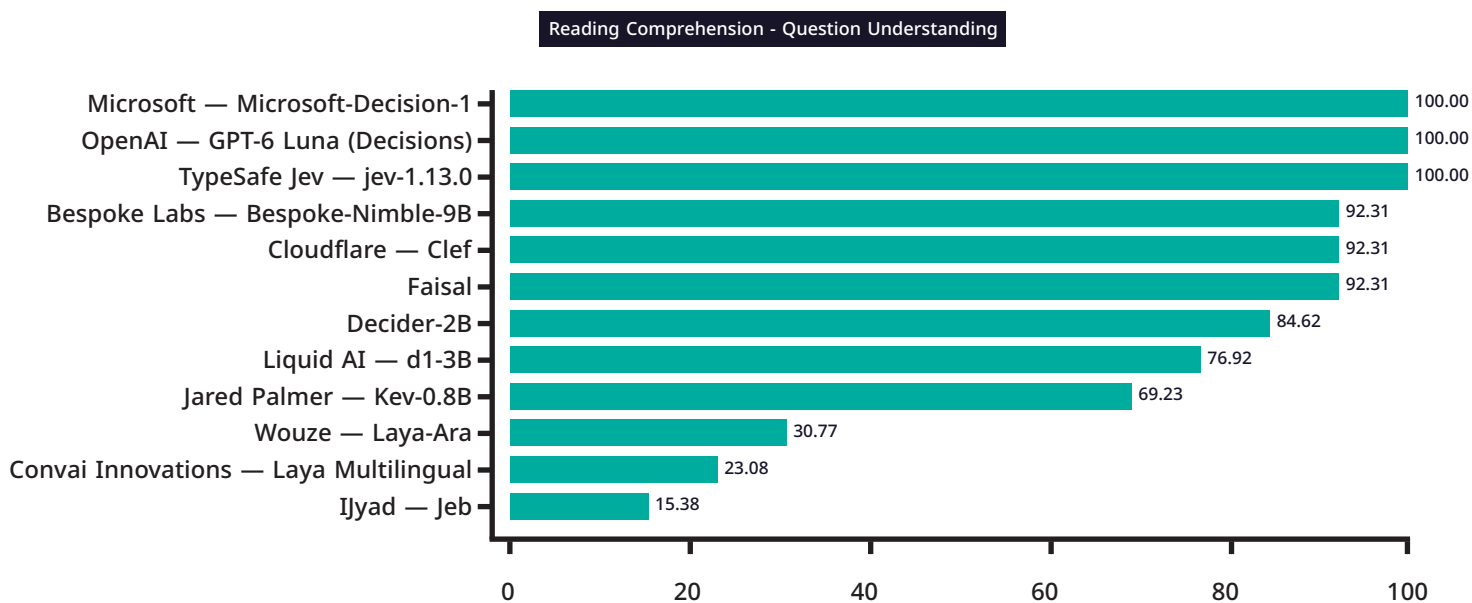


Subtasks:

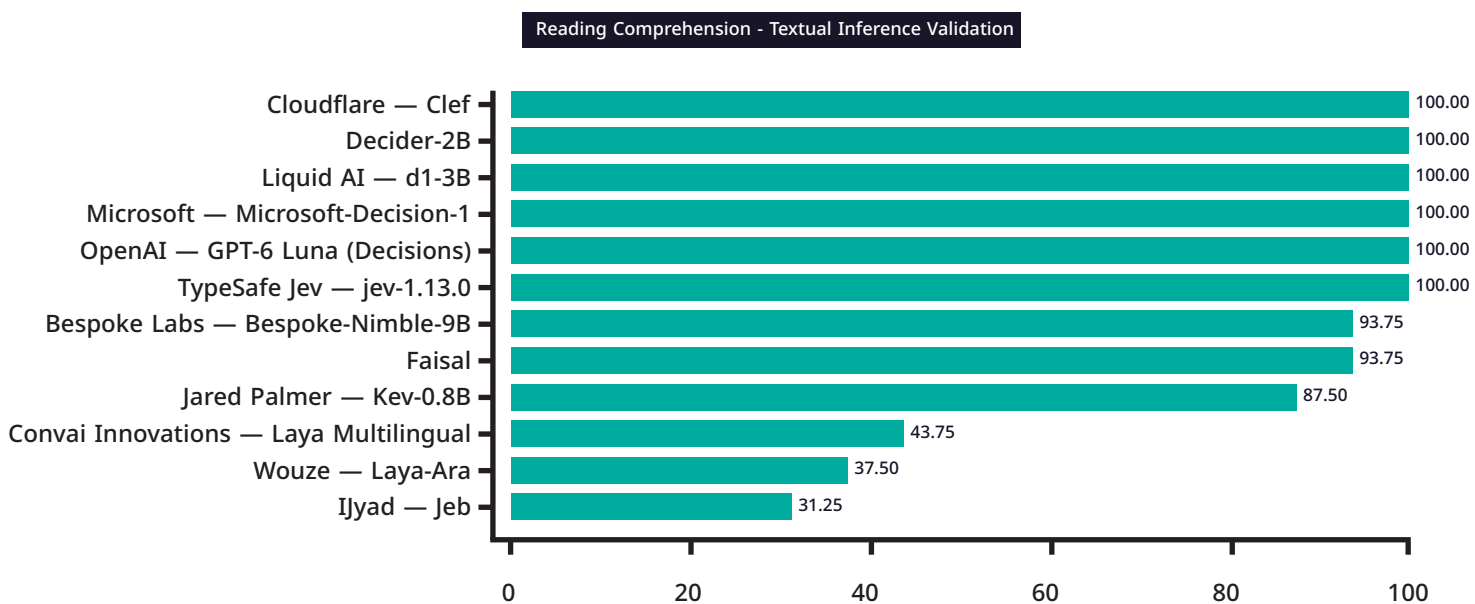
9.1 Answerability Classification



9.2 Question Understanding



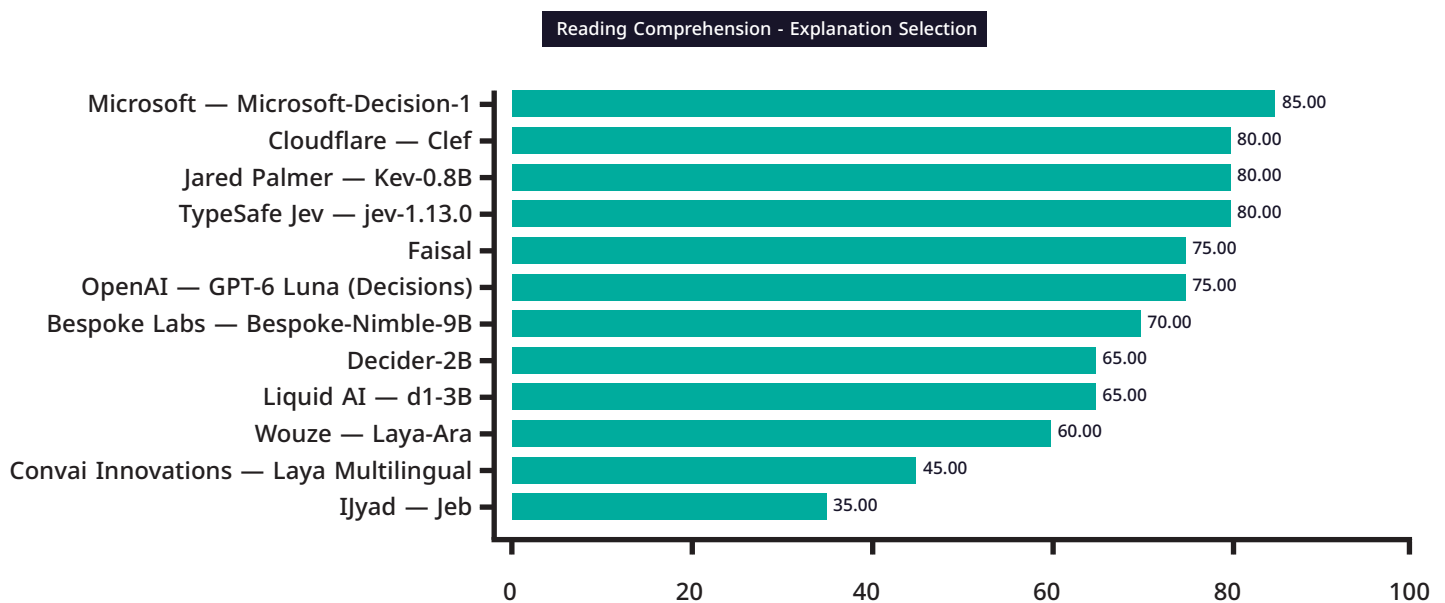
9.3 Textual Inference Validation



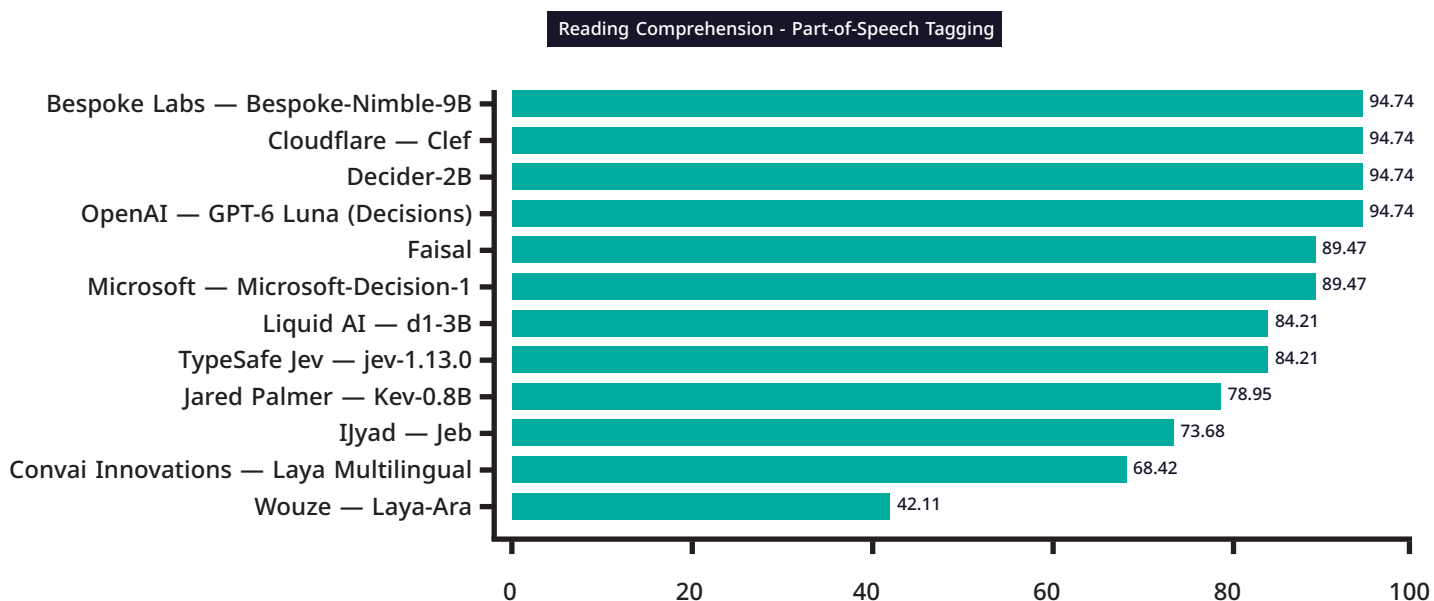
9.4 Dialogue Selection



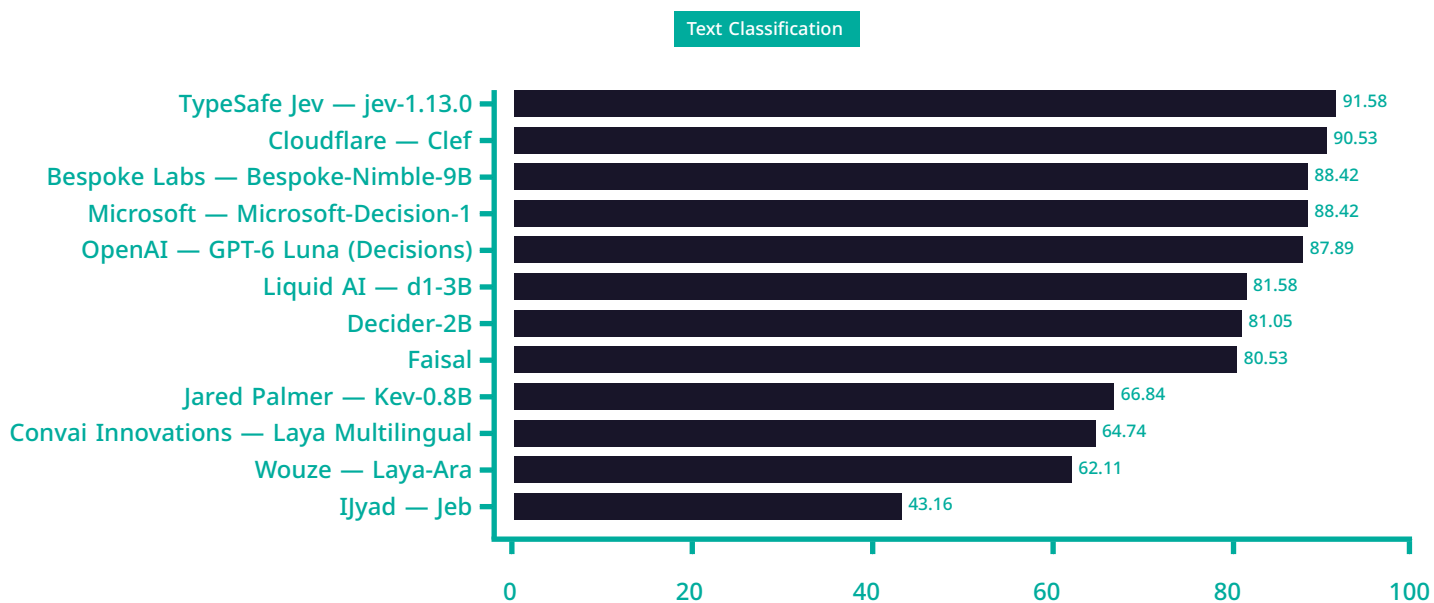
9.5 Explanation Selection



9.6 Part-of-Speech Tagging

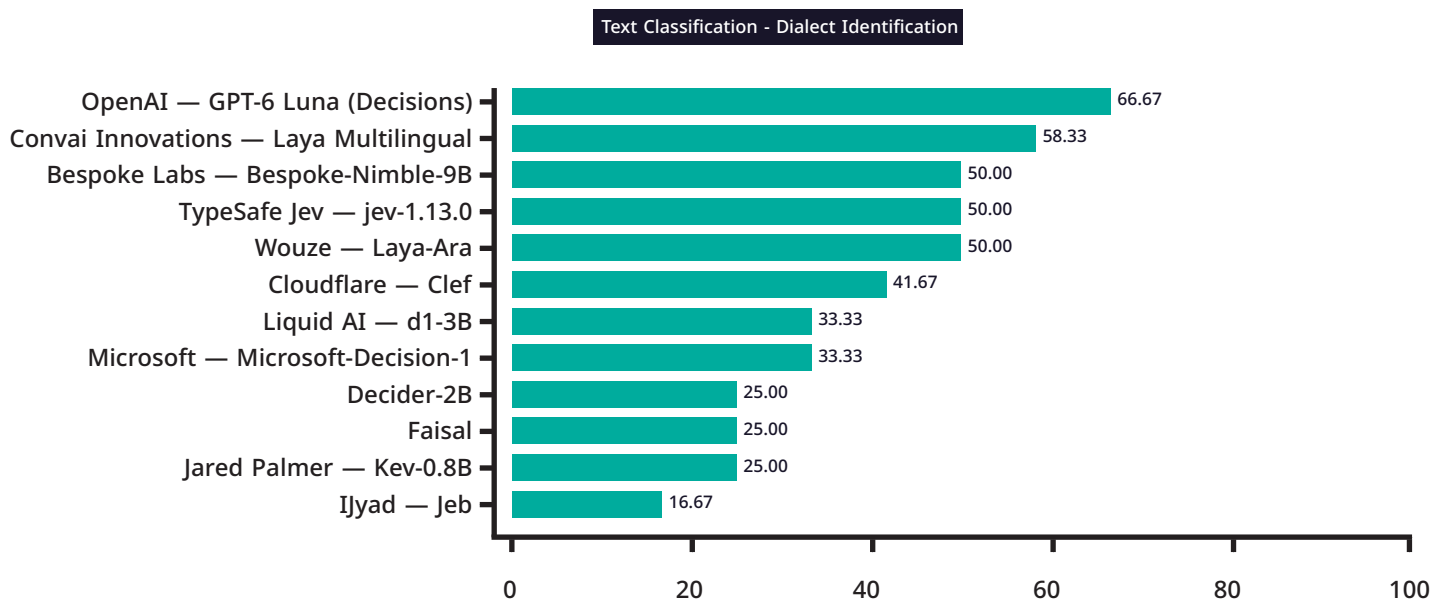


10. Text Classification

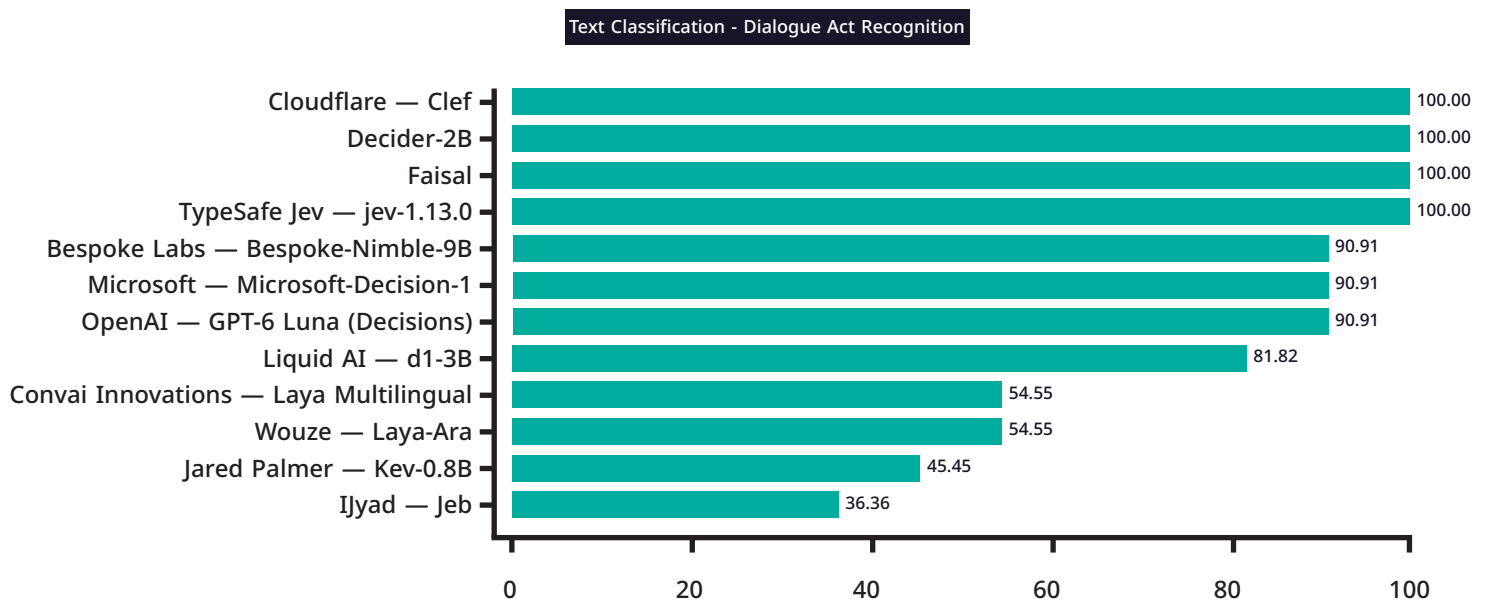


Subtasks:

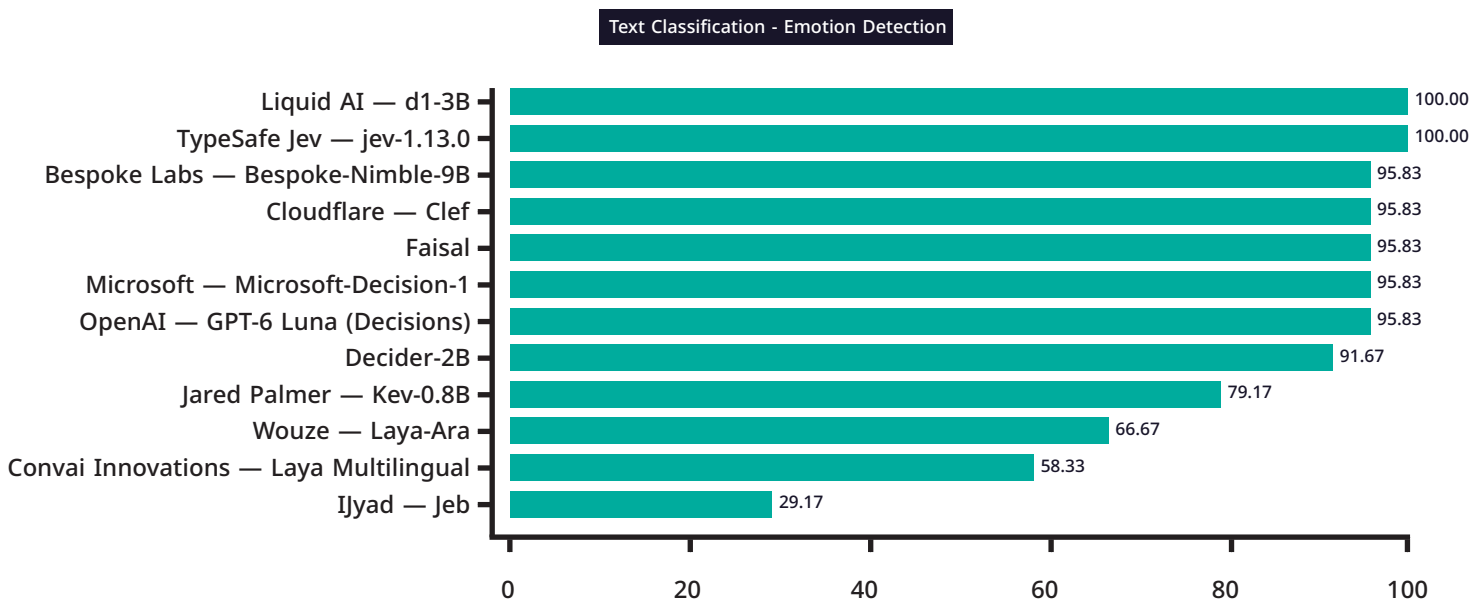
10.1 Dialect Identification



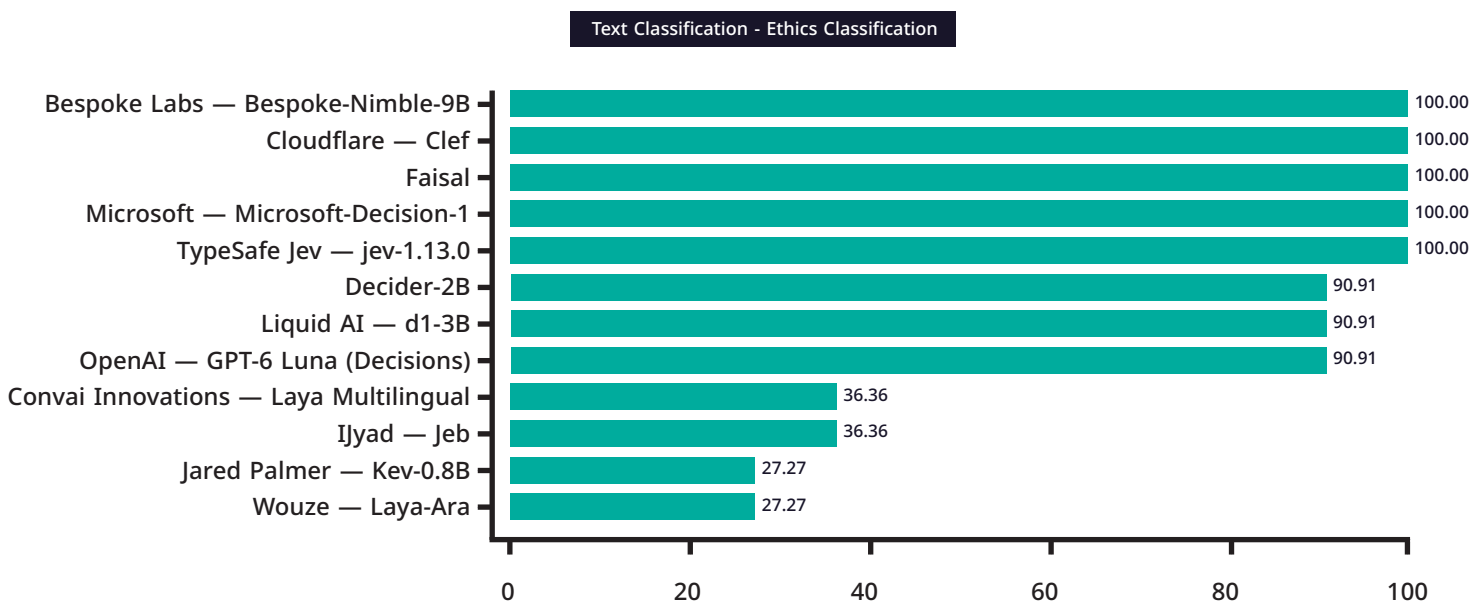
10.2 Dialogue Act Recognition



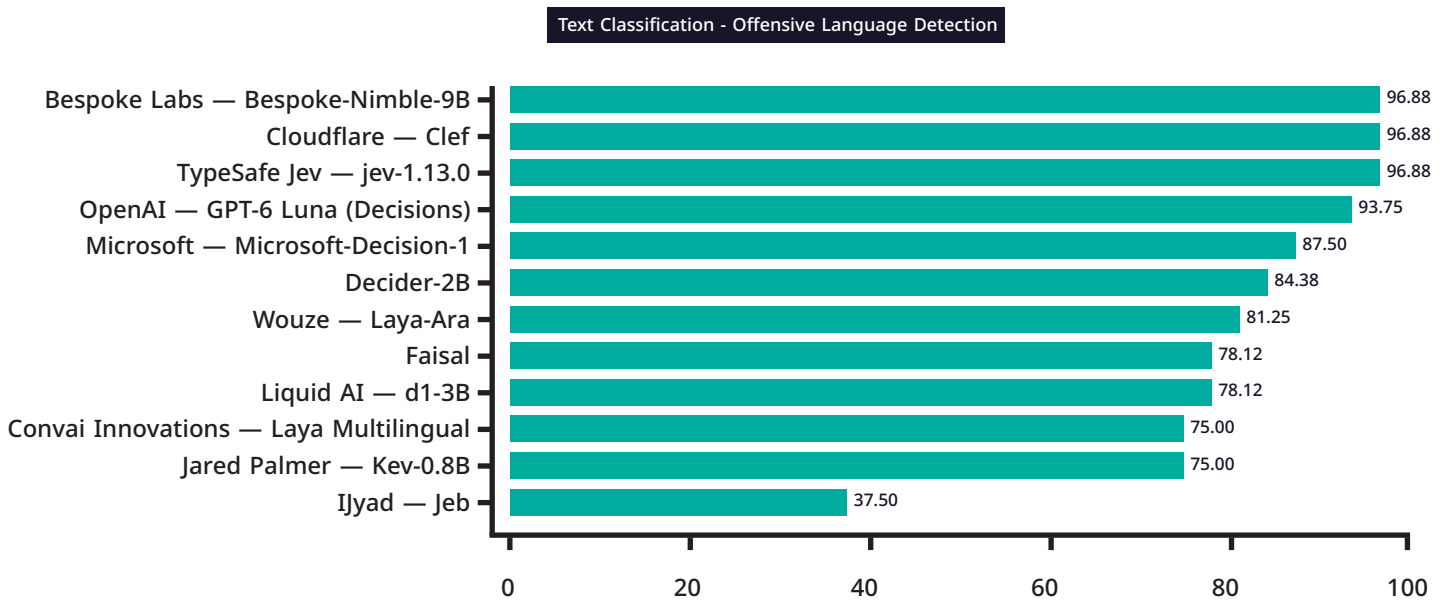
10.3 Emotion Detection



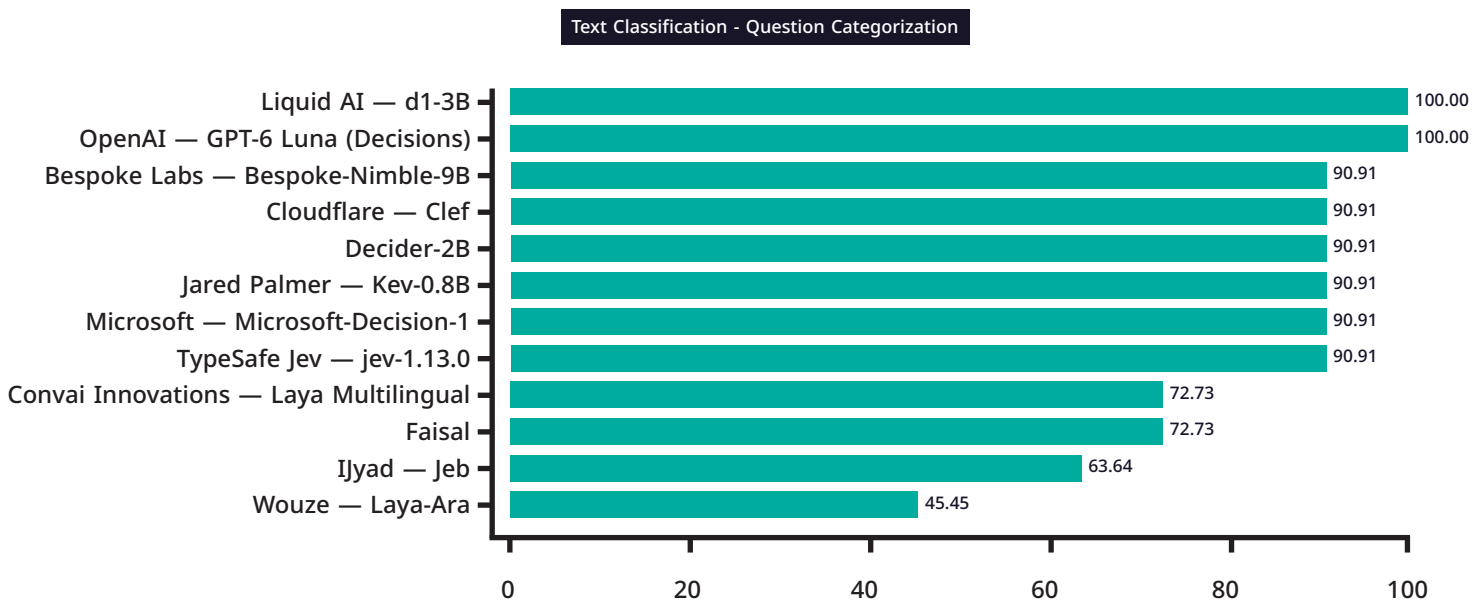
10.4 Ethics Classification



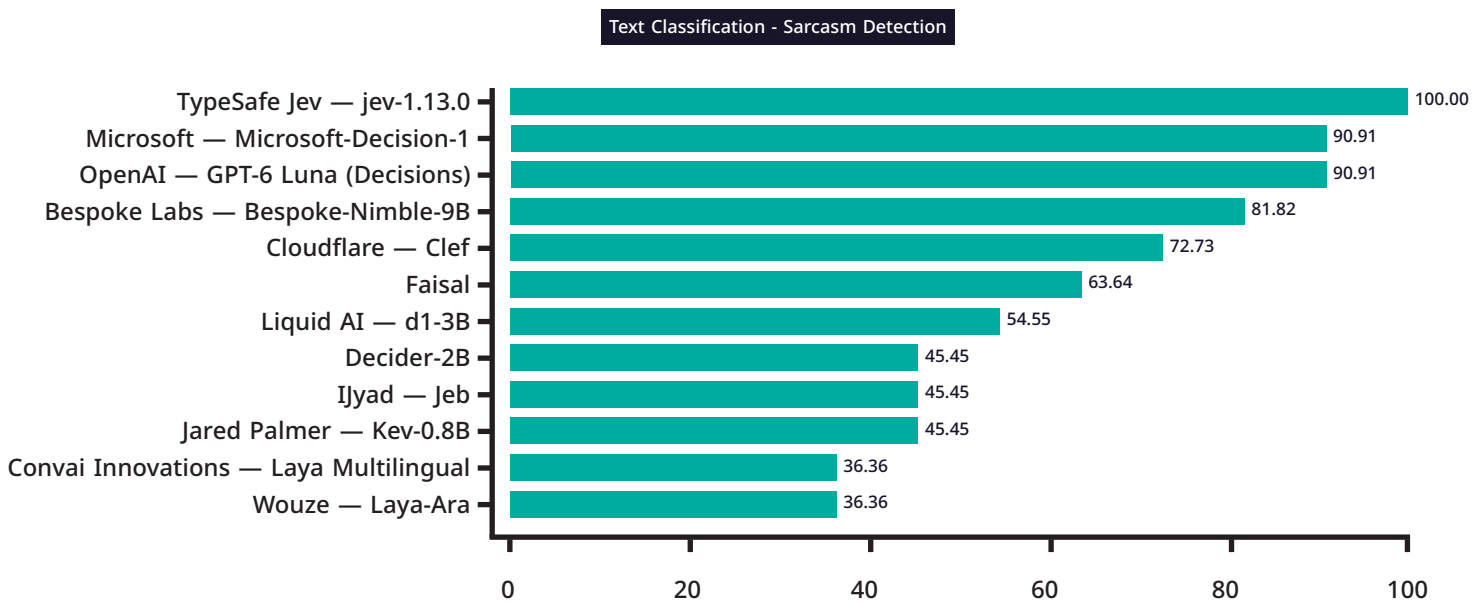
10.5 Offensive Language Detection



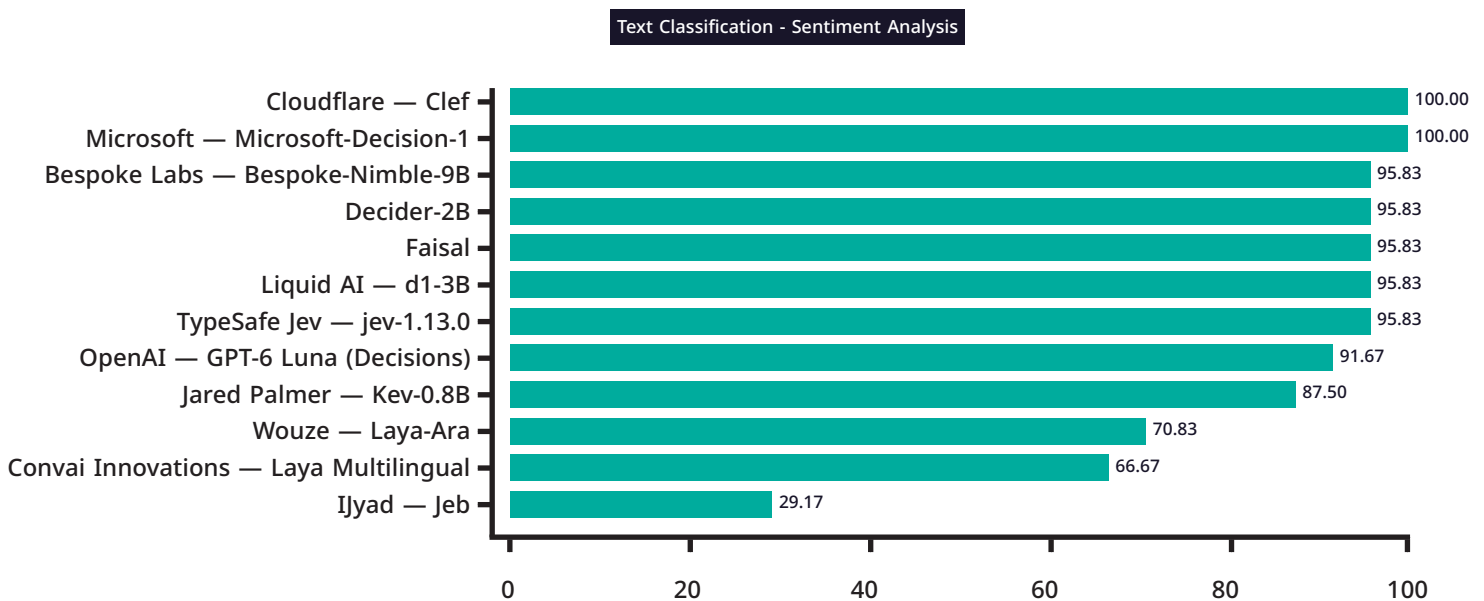
10.6 Question Categorization



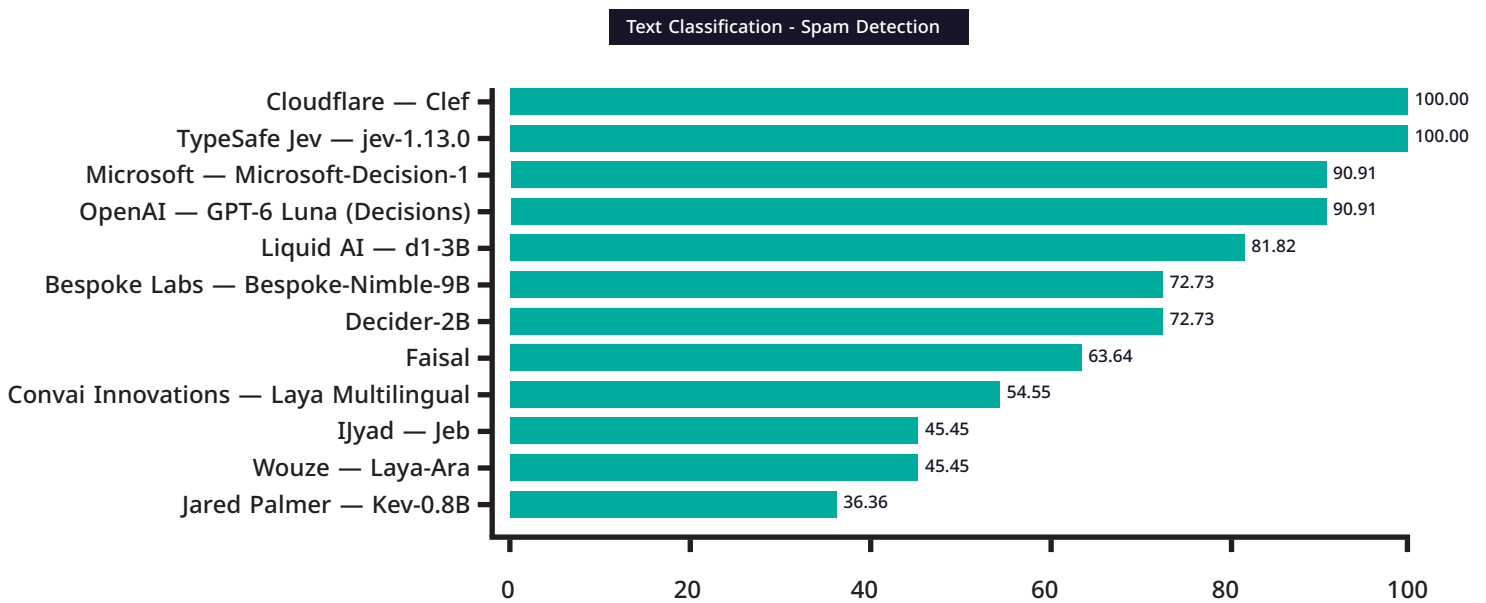
10.7 Sarcasm Detection



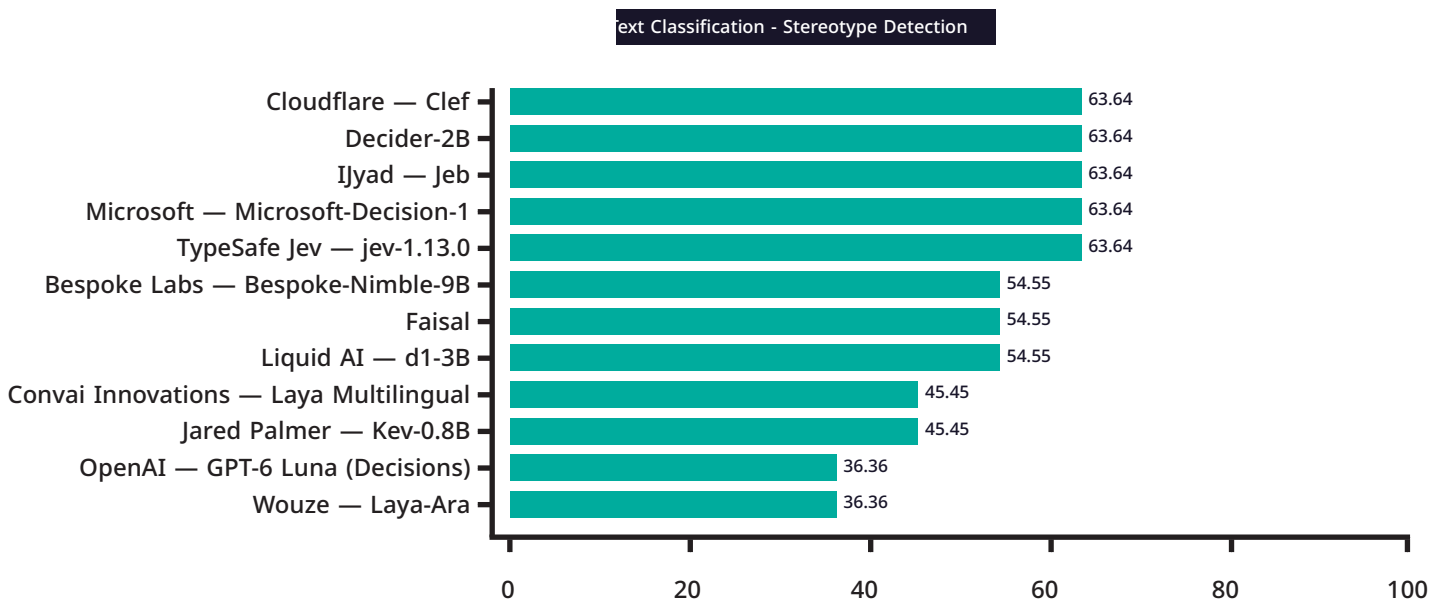
10.8 Sentiment Analysis



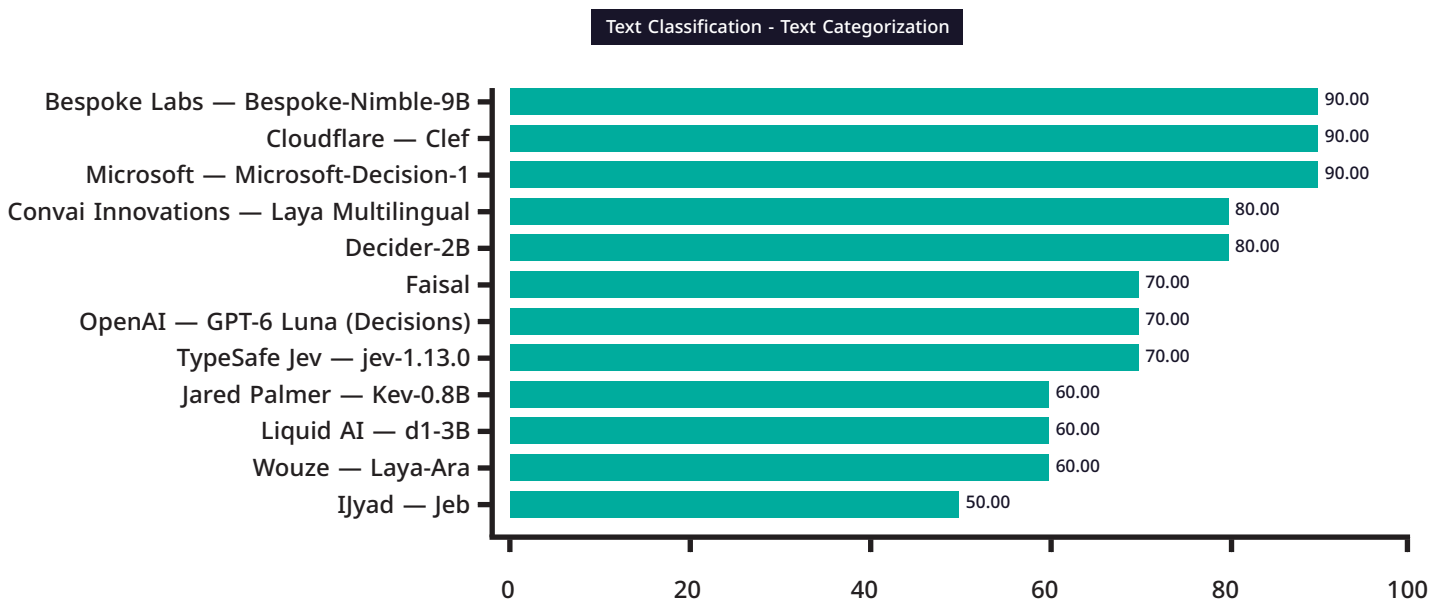
10.9 Spam Detection



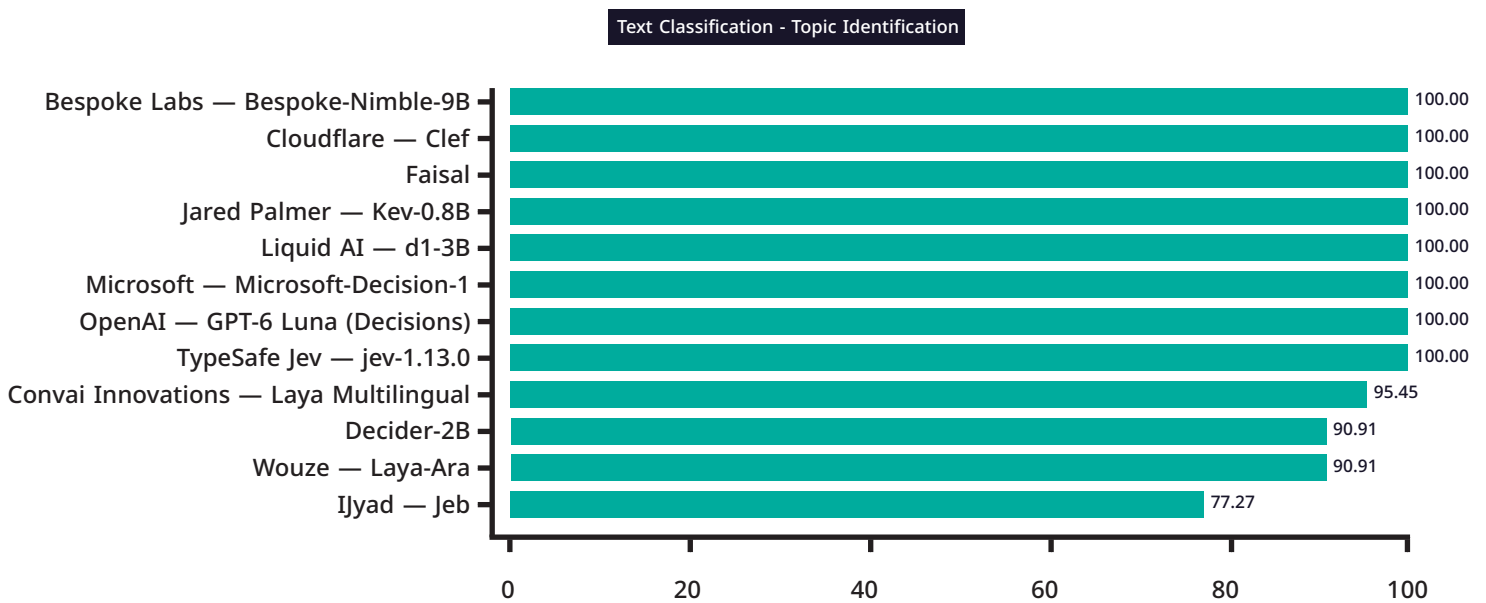
10.10 Stereotype Detection



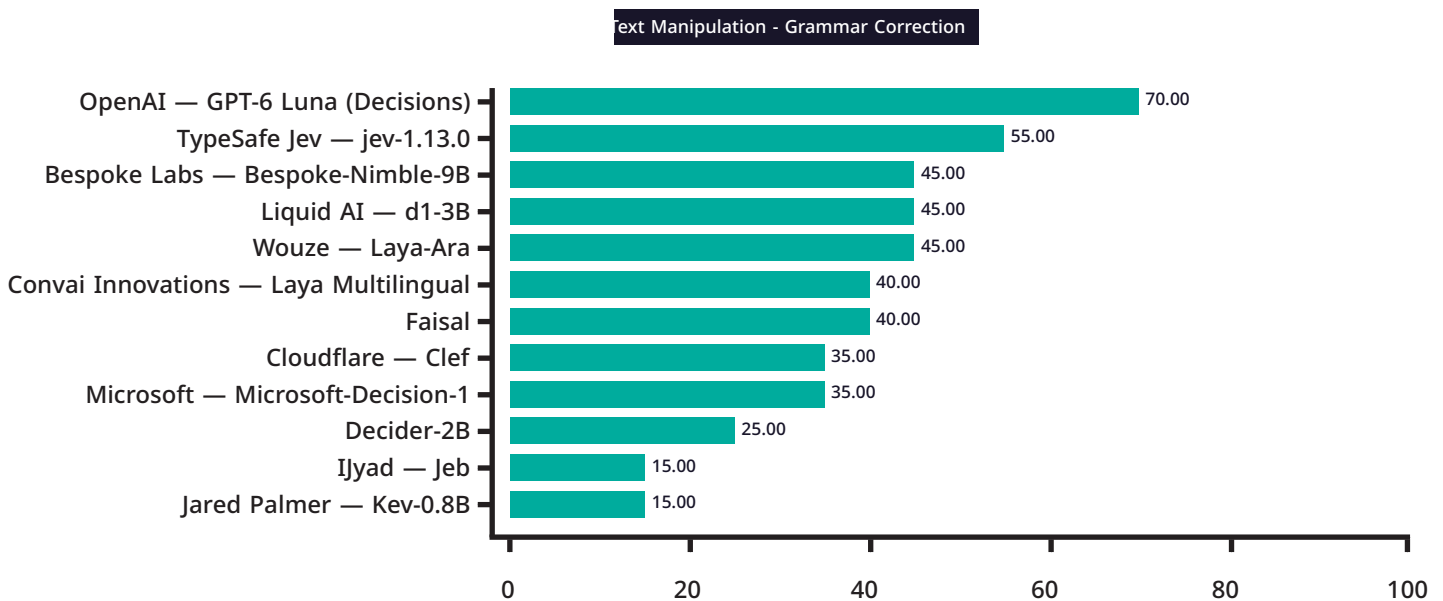
10.11 Text Categorization



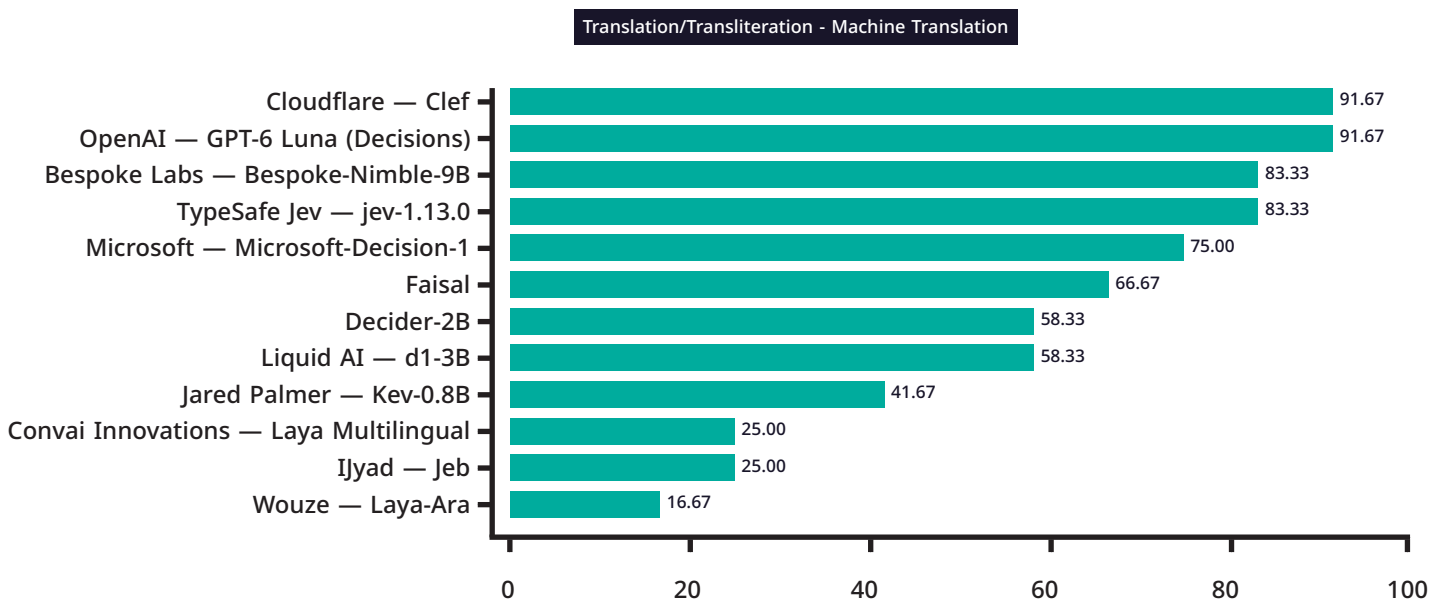
10.12 Topic Identification



11. Grammar Correction



12. Translation/Transliteration



✓ Appendix (A): Definitions of Categories and Tasks

Category	Category Definition	Task	Task Definition
Creative Writing	Tasks related to producing creative text and assessing how well a continuation fits it.	Text Completion	Choosing the most suitable continuation for an incomplete text.
		Text Continuation Evaluation	Judging how suitable a proposed continuation of the text is.
Entailment	Measuring the semantic relationship between two texts: similarity in meaning or whether one entails the other.	Semantic Similarity	Estimating the degree of closeness in meaning between two texts.
		Textual Entailment	Determining whether a text entails, contradicts, or is neutral toward another.
Factuality	Verifying the correctness of an answer or a claim based on reference information.	Answer Verification	Judging the correctness of an answer against its question or reference.
		Claim Verification	Judging the correctness of a claim based on information.
Fill in the Blank	Choosing the appropriate word or phrase to complete a blank within the text.	Fill in The Blank	Choosing what correctly completes the blank in the sentence.
Information Extraction	Identifying entities, relations, and required information within the text.	Entity Categorization	Determining the type or category of an entity mentioned in the text.
		Entity Relation Classification	Determining the type of relation linking two entities.
		Extracting Required Information	Selecting a specific piece of required information from the text.

Category	Category Definition	Task	Task Definition
Logic	Logical reasoning, textual coherence, cause-and-effect relations, and sentence ordering.	Cause Effect Classification	Distinguishing the cause from the effect between two events.
		Coherence Classification	Judging how connected and coherent the text is.
		Commonsense Validation	Distinguishing a logical statement from one that defies common sense.
		Logical Reasoning	Inferring the correct conclusion from given premises.
		Sentence Ordering	Choosing the correct order for shuffled sentences.
Program Execution	Inferring the result of executing programming steps or instructions.	Program Execution	Inferring the output of executing programming instructions.
Question Answering	Choosing the correct answer to a posed question.	Answering Given Question	Choosing the correct answer to a given question.
Reading Comprehension	Understanding the text: answerability, question understanding, inference, and analysis of its components.	Answerability Classification	Determining whether the text makes it possible to answer the question.
		Dialogue	Choosing the most suitable reply or conversational turn.
		Explanation	Choosing the most suitable explanation for a phenomenon or text.
		Question	Grasping the intent of the question and identifying its correct interpretation.
		Textual Inference Validation	Judging the correctness of an inference drawn from the text.
		Part-of-Speech Tagging	Determining the morphological type of the word (noun/verb/particle...).

Category	Category Definition	Task	Task Definition
Text Classification	Classifying text according to a semantic category	Dialect	Distinguishing the Arabic dialect of the text.
		Dialogue Act Recognition	Classifying the function of an utterance in a dialogue (question, request, ...).
		Emotion Detection	Identifying the emotion expressed in the text.
		Ethics Classification	Judging the ethical acceptability of a behavior or statement.
		Offensive Language Detection	Determining the presence of offensive language in the text.
		Question Categorization	Classifying the question according to its type or domain.
		Sarcasm Detection	Distinguishing sarcastic text from non-sarcastic text.
		Sentiment Analysis	Determining the affective orientation (positive/negative/neutral).
		Spam Detection	Distinguishing a spam message from a legitimate one.
		Stereotype Detection	Determining the presence of a stereotype in the text.
		Text Categorization	Assigning the text to a topical category.
		Topic Identification	Identifying the main topic of the text.

Category	Category Definition	Task	Task Definition
Text Manipulation	Detecting grammatical errors in the text and correcting them.	Grammar Correction	Detecting the grammatical error and choosing the correct fix.
Translation/Transliteration	Translating text between languages.	Machine Translation	Choosing the correct translation of the text between the two languages.



مجمع الملك سلمان
العالمي للغة العربية
King Salman Global Academy for Arabic Language



برنامج تنمية
القدرات البشرية
Human Capability
Development Program



رؤية VISION
2030
المملكة العربية السعودية
KINGDOM OF SAUDI ARABIA

