signal65

AGENTIC AI CAPABILITIES TESTING Q1 2026

Benchmarking Leadership in Open and Proprietary Models

Evaluating Agentic AI Capabilities with the KAMI v0.1 Benchmark

AUTHOR

Mitch Lewis
Performance Analyst | Signal65

JANUARY 2026

Executive Summary

This report evaluates agentic AI capability with the Kamiwaza Agentic Merit Index (KAMI) Benchmark and provides analysis of the agentic AI landscape as of Q1 2026. The KAMI Benchmark provides measurement of AI model accuracy during enterprise-focused agentic workloads. This testing expands upon previous results from the KAMI v0.1 benchmark, outlined in the Signal65 report Measured Leadership with Agentic AI on Open Models. While the previous report focused primarily on popular open source LLMs, additional testing in this Q1 update has broadened the test set and includes several prominent proprietary models.

Key Highlights

- GPT-5 is the top agentic AI performer at **95.7% mean accuracy score**
- **GLM-4.6 leads all open models** with 92.57% mean accuracy
- Open models achieve **7 of the top 10** highest accuracies for agentic workloads

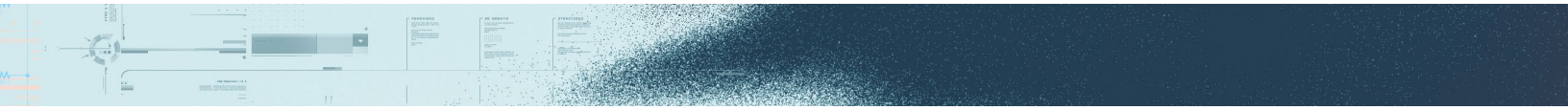
Key findings include:

- GPT-5 leads all models tested with a mean accuracy of 95.7%
- Top open source models are highly competitive with proprietary models, in many cases out-performing leading proprietary models.
  - GLM-4.6 achieves the highest overall score for an open source model, with 92.57% mean accuracy.
  - DeepSeek-v3.1 achieved the second highest overall score for an open source model with 92.19% accuracy.
  - Qwen3-Coder-480B-A35B-Instruct achieved the third highest overall score for an open source model with 91.88% accuracy.
- Qwen3-Next-80B-A3B-Instruct achieved the highest score of any open model with fewer than 100B parameters at 83.79% accuracy.
- Ongoing model development of proprietary models shows inconsistent improvement for agentic use cases, with some newer models underperforming previous generations.
  - GPT-5 notably outperforms newer GPT models, including GPT-5.1 and GPT-5.2
  - Claude-Haiku-3.5 significantly outperforms Claude-Haiku-4.5.
  - Similar discrepancies are seen in open model families, including Qwen, Llama, and MiniMax.
- Some models achieved higher accuracy when run on AWS Bedrock than on on-premises hardware, indicating that infrastructure and configuration can impact agentic accuracy.

An overview of the top 10 performing models tested can be seen below:

Rank	Model	Mean Accuracy Score
1	GPT-5 (Medium Reasoning)	95.7%
2	GLM-4.6	92.57%
3	DeepSeek-v3.1	92.19%
4	Qwen3-Coder-480B-A35B-Instruct	91.88%
5	Qwen3-235B-A22B-Instruct-2507	90.37%
6	MiniMax-M2	89.89%
7	Claude-Sonnet-4.5	89.63%
8	GPT-5.2 (Medium Reasoning)	89.08%
9	Qwen3-235B-A22B-Instruct-2507-FP8	88.75%
10	GLM-4.5	88.14%

Figure 1: *KAMI v0.1 Benchmark Top 10 Results (Q1 2026)*



About the KAMI Benchmark

The Kamiwaza Agentic Merit Index (KAMI) provides a unique AI benchmark targeted at understanding LLM performance in real-world, agentic AI scenarios. The KAMI benchmark, which is a joint collaboration between Signal65 and Kamiwaza, was developed to address key challenges experienced with other AI benchmarks and to establish a benchmark that accurately represents real enterprise agentic AI workloads. Key features of the KAMI benchmark include:

- **Multi-level Randomization** – To prevent memorization of questions and answers, which can easily be consumed through a model's training data, KAMI utilizes multi-level randomization to create a dynamic, non-memoizable benchmark. Each question or task asked of an LLM is based on a static prompt that is dynamically augmented with randomized variables. The sandbox environment that the agent can interact with – including files, directory structures, and databases – is additionally randomized. This enables a repeatable benchmark that ensures models cannot simply memorize the correct results.
- **Deterministic Scoring** – Although each run of the KAMI benchmark randomizes the data and test environment, accuracy of the benchmark is maintained with deterministic scoring. All answers for each unique test run are generated at runtime, enabling real ground-truth scoring and avoiding complexities of LLM judges.

- **Agentic AI Focus** - The KAMI benchmark is purpose built to evaluate agentic AI tasks that may commonly be seen in real world enterprise environments. The benchmark provides LLMs access to tools that can be used to complete various tasks, such as creating files or accessing databases. This enables LLMs to complete loops of inference and tool calling, providing insight into real agentic capabilities rather than single-shot question and answering.

An example of a randomized agentic task included in the KAMI benchmark is provided in the figures below. This question requires models to query information from a database to answer a specific business question.

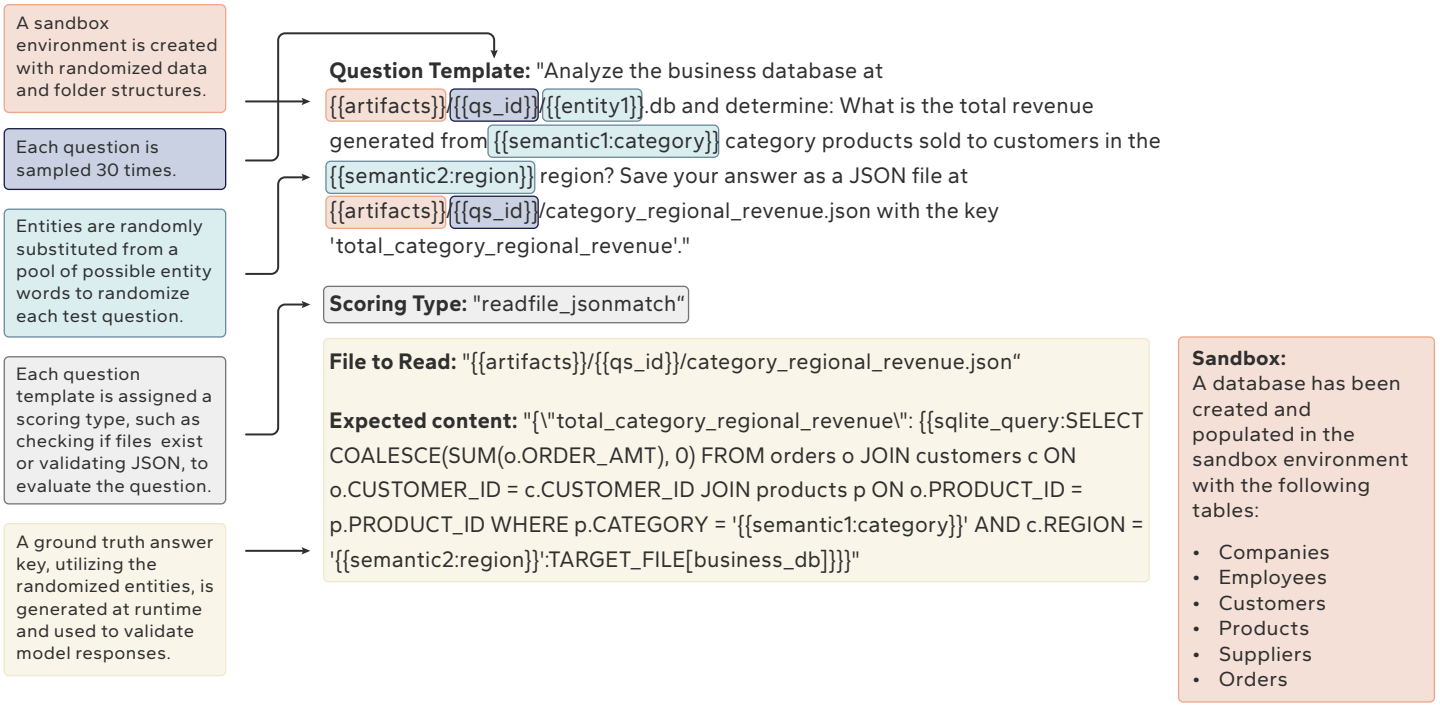


Figure 2: *Database Question Template Overview*

Example:

"Analyze the business database at test_artifacts/q503_s11/harbor.db and determine: What is the total revenue generated from technology category products sold to customers in the west region? Save your answer as a JSON file at test_artifacts/q503_s11/category_regional_revenue.json with the key 'total_category_regional_revenue'."

Correct Answer:

File: test_artifacts/q503_s11/category_regional_revenue.json
Contents: {"total_category_regional_revenue": 10000}

The correct answer correctly retrieved the answer to the query, formatted it as JSON, and created the output file in the correct directory.

Incorrect:

File: test_artifacts/q503_s11/category_regional_revenue.json
Contents: {"total_category_regional_revenue": 500}

Any answer that doesn't match the expected answer key is incorrect. In this example, the agent correctly created a JSON file with correct formatting, but incorrectly queried the database to calculate the value for 'total_category_regional_value'.

Figure 3: *Database Question Example*

More detailed information on the design, characteristics, and unique features of the KAMI benchmark can be found in previous reports from both Signal65 and Kamiwaza.

Test Overview

The KAMI v0.1 Benchmark contains 19 distinct question templates, grouped into 7 specific categories. All questions were sampled 30 times for each run of the KAMI test suite to accommodate the variance of the randomized questions. In addition, for each model tested, the entire test suite was run multiple times and models were scored using their mean accuracy over all runs. An overview of the test questions can be seen in Figure 4 below.

Category	Test Summary
Basic Reasoning	Respond only with a specific word.
	Respond with multiple specified words in a specified order.
File System Operations	Create specific files in a specified directory.
	Create specific directory structures and include various files.
Text Search and Extraction	Find two specific lines from a file.
	Find several specific lines from an extended file.
	Retrieve two specific words from a text file.
	Retrieve several specific words from an extended text file.
CSV Processing	Create JSON summary of a CSV file.
	Analyze business data across multiple CSV files. Answer 6 specific questions.
	Analyze business data across multiple CSV Files. Single question.
Database Processing	Query business database to fine number of orders over a specified value within a specified region.
	Analyze business database and create a comprehensive report. 6 specific questions.
	Analyze business database to find total revenue from a specified product in a specified region.
Database Processing (Guided)	Repeat simple database task with a hint given.
	Repeat complex database task with a hint given.
Response Format Instruction Following	Output answer to txt file.
	Output answer in JSON format.
	Output number only.

Figure 4: *KAMI Question Overview*

This testing follows the same processes as in the first KAMI v0.1 Benchmark report, with an expanded test set. The first iteration of KAMI testing ran the KAMI v0.1 test suite on 31 models, with a focus on open source models. The original test set included one proprietary model – Claude-3.5-Haiku-20241022 - as a single comparison point. This testing expands upon the previous data set with an additional 39 models, including more open source models and several notable proprietary models. An overview of all models tested can be seen in Figure 5.

Model Family	Models	
Amazon Nova	• Nova-Premier • Nova-Pro • Nova-Lite	• Nova-2-Lite • Nova-Micro
Anthropic Claude	• Claude-Sonnet-4.5 • Claude-3.5-Haiku-20241022	• Claude-Haiku-3.5 • Claude-Haiku-4.5
DeepSeek	• DeepSeek-V3.1	• DeepSeek-V3
Google Gemini	• Gemini-3-Pro-Preview • Gemini-2.5-Pro • Gemini-2.5-Flash	• Gemini-2.5-Flash-Lite • Gemini-2.0-Flash • Gemini-2.0-Flash-Lite
IBM Granite	• Granite-4.0-H-Small • Granite-4.0-H-Tiny	• Granite-4.0-H-Micro
Kimi	• Kimi-K2-Thinking	
Meta Llama	• Llama-4-Maverick-17B-128E-Instruct • Llama-4-Maverick-17B-128E-Instruct-FP8 • Llama-4-Scout-17B-16E-Instruct • Llama-3.3-70B-Instruct	• Llama-3.3-70B-Instruct-FP8-KV • Llama-3.1-70B-Instruct • Llama-3.1-8B-Instruct
Microsoft Phi	• Phi-4	
MiniMax	• MiniMax-M2.1	• MiniMax-M2
Mistral	• Mistral-Large-Instruct-2411	• Mistral-Large-3-675B-Instruct-2512
OpenAI GPT	• GPT-5.2 (Medium Reasoning) • GPT-5.1 • GPT-5.1 (Medium Reasoning) • GPT-5 (Medium Reasoning)	• GPT-5-Mini (Medium Reasoning) • GPT-5-Nano (Medium Reasoning) • GPT-4.1
Qwen	• Qwen3-Coder-480B-A35B-Instruct • Qwen3-235B-A22B-Instruct-2507-FP8 • Qwen3-235B-A22B-Instruct-2507 • Qwen3-Max-2025-09-23 • Qwen-Plus-2025-09-11 • Qwen3-Next-80B-A3B-Instruct • Qwen3-Max-Preview • Qwen3-30B-A3B (Thinking Mode) • Qwen3.30B-A3B-Instruct-2507 • Qwen3-30B-A3B • Qwen3-14B (Thinking Mode) • Qwen3-8B (Thinking Mode) • Qwen-Flash-2025-07-28 • Qwen3-235B-A22B	• Qwen3-32B (Thinking Mode) • Qwen3-32B-FP8 • Qwen3-32B • Qwen3-14B-FP8 • Qwen3-14B • Qwen3-8B • Qwen3-4B-Instruct-2507 • Qwen3-4B (Thinking Mode) • Qwen3-4B • Qwen2.5-72B-Instruct • Qwen2.5-32B-Instruct • Qwen2.5-14B-Instruct • Qwen2.5-7B-Instruct
Z.ai GLM	• GLM-4.6 • GLM-4.5	• GLM-4.5-Air

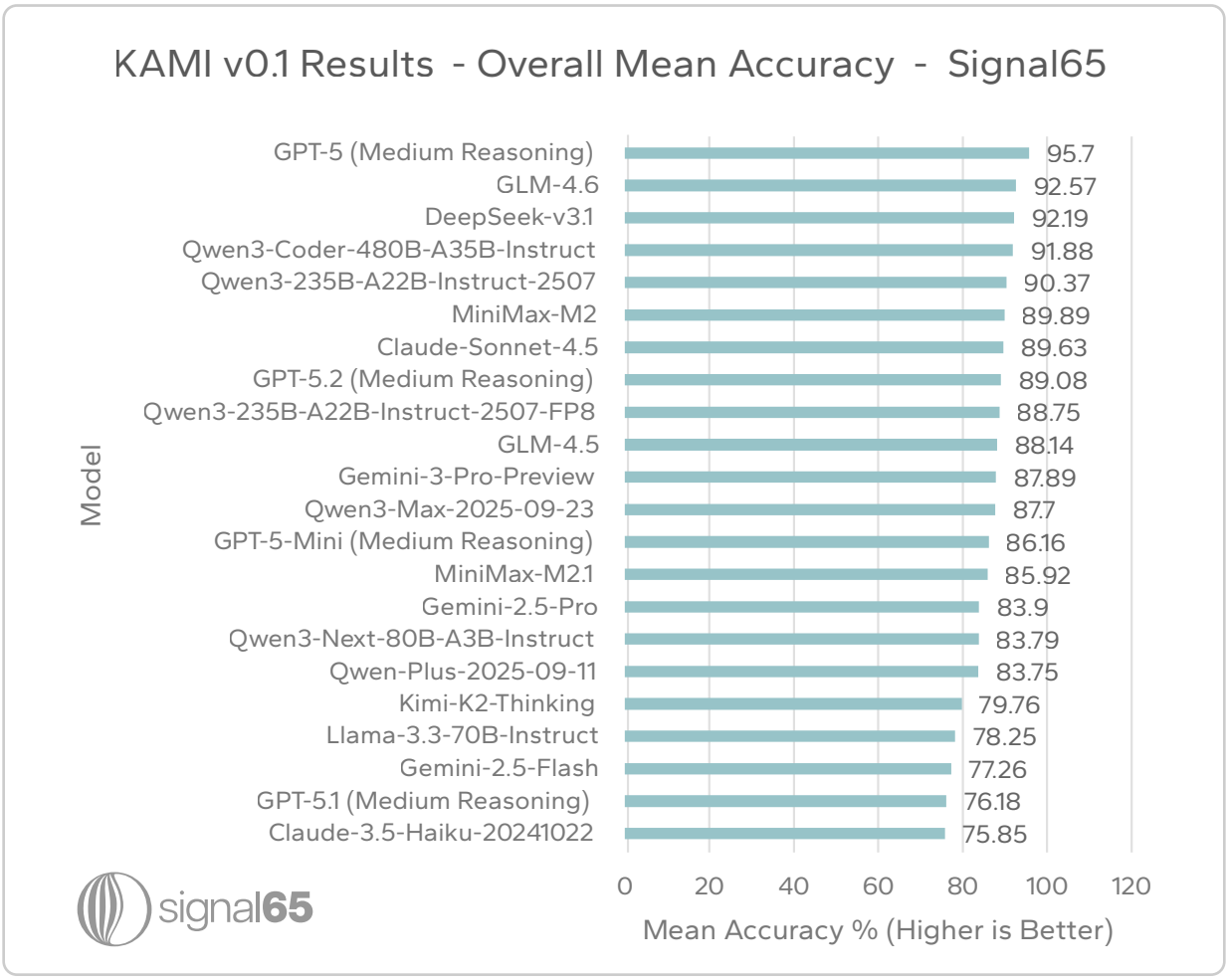
Figure 5: *Models Tested*

Signal65 Comment *– Models, such as certain Qwen and GPT models, can be configured with varying levels of thinking or reasoning capabilities. Models run with these capabilities have been labeled to distinguish their configurations.*

Models were run across a range of infrastructure, including hardware in the Signal65 AI Lab, proprietary model API endpoints, and AWS Bedrock. This benchmark provides a measurement of model accuracy, not hardware performance – however, to ensure consistency, some models were tested on multiple hardware platforms. To fairly represent each model, the highest score for each model has been selected, regardless of hardware platform. For most models, variance between platforms has been found to be statistically insignificant, and attributable to randomization within the test set. In a few scenarios, models run on AWS Bedrock were found to outperform on-premises deployments, these models will be discussed in more depth in the following results.

Results

An overview of the full results can be seen in Figure 6.



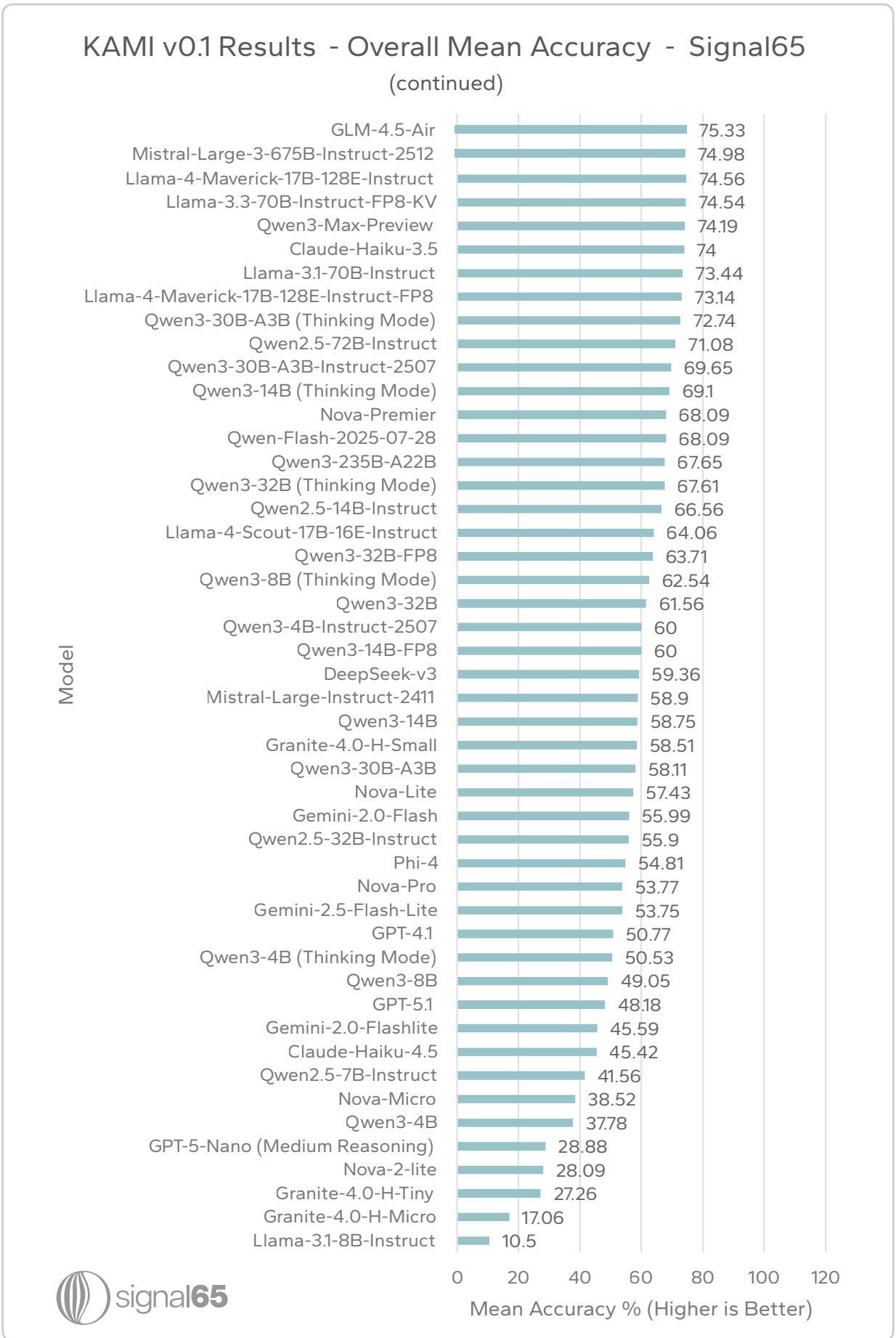


Figure 6: *KAMI v0.1 Results Overview*

Notably, these results present new leading models compared to the previous iteration of KAMI v0.1 testing. In the first set of models tested, Qwen3-235B-A22B-Instruct-2507-FP8 achieved the highest overall mean accuracy at 88.75%. With additional testing, eight of the top ten models have surpassed this score:

- GPT-5 (Medium Reasoning)
- GLM-4.6, DeepSeek-v3.1
- DeepSeek-v3.1
- Qwen3-Coder-480B-A35B-Instruct
- Qwen3-235B-A22B-Instruct-2507
- MiniMax-M2
- Claude-Sonnet-4.5
- GPT-5.2 (Medium Reasoning)

Five of these models surpassed 90% mean accuracy, which was not achieved by any of the models in the first round of testing. The leading model, GPT-5 recorded a particularly impressive score with 95.7% mean accuracy.

***Signal65 Comment** – A notable change in this second batch of test results is the improvement Qwen3-235B-A22B-Instruct-2507. This model was included in the initial test set and achieved the second highest result, at 88.4% mean overall accuracy, second only to its own FP8 variation. During ongoing testing, this model was re-run on AWS Bedrock, achieving a notably higher score of 90.37%, which was included according to the test methodology of retaining each model's highest score. This model is one example of models achieving a statistically significant accuracy when run on AWS Bedrock.*

The top 10 models show an interesting mix of proprietary and open source models. While a proprietary model, GPT-5, achieved the highest overall score, only two other proprietary models, GPT-5.2 and Claude-Sonnet-4.5 ranked in the top 10. All other models in the top 10 are open source, led by GLM-4.6. This indicates that while there may be some advantages for certain proprietary models, there is not a broad gap between proprietary and open source models.

Basic Reasoning Tasks

The basic reasoning tasks included in the KAMI v0.1 Benchmark exist as simple evaluation of a model's capability to perform basic tasks when given tool access. Models that are challenged with these tasks are likely not well suited for agentic use cases, as tool access impacts basic functionality such as returning a specific word. In total 44 of the 65 models tested achieved 100% accuracy for these tasks. Five models scored below 90%.

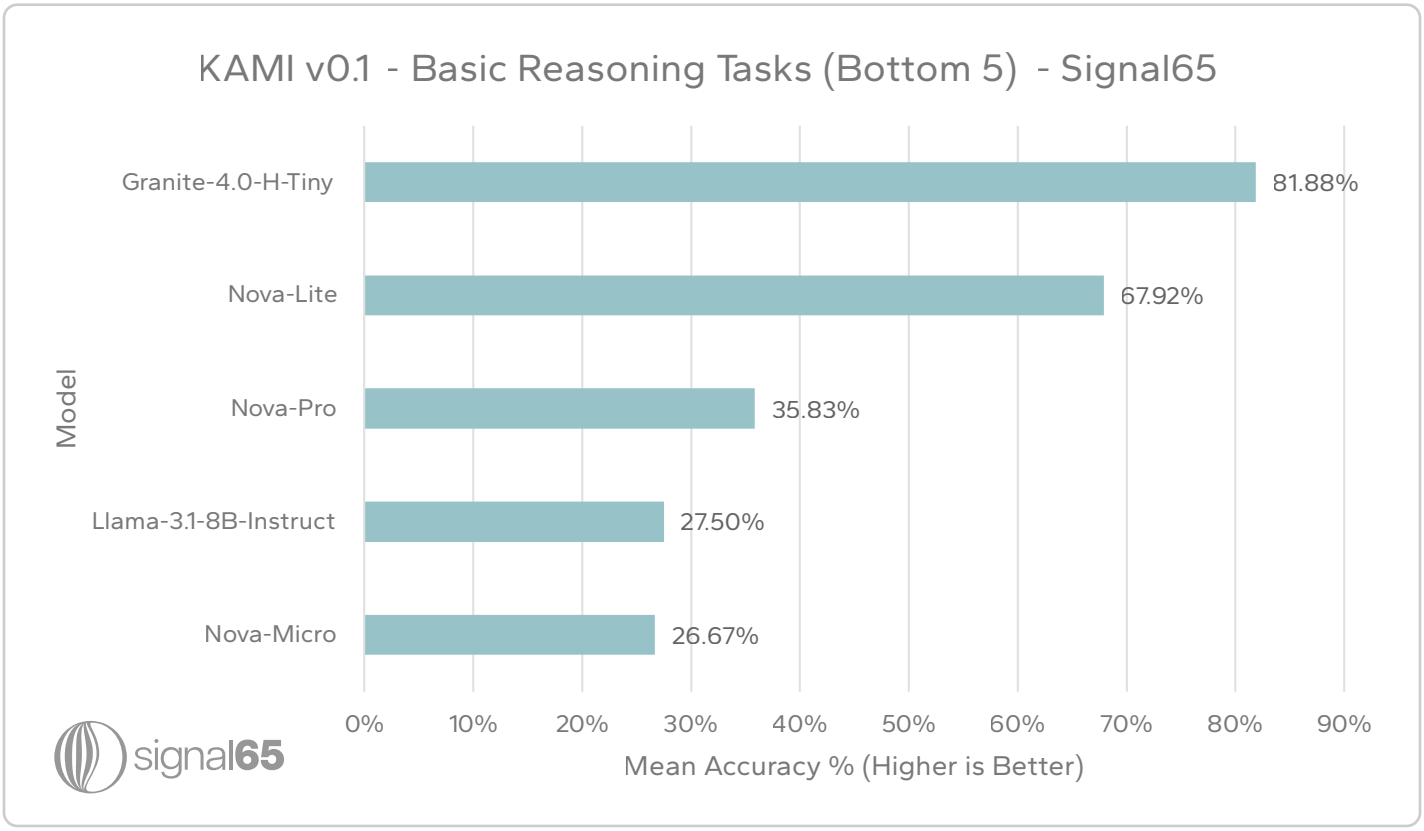


Figure 7: *KAMI v0.1 Basic Reasoning Tasks (Bottom 5)*

Filesystem Operations

In general, the first round of KAMI testing found most models to be highly successful across the Filesystem Operations tasks, with six of the 31 models achieving 100% accuracy, and a majority of models achieving over 90% accuracy. In the expanded test set, the number of models achieving 100% accuracy rose to 11 out of 70, with an additional 20 scoring 99% or above. Models that achieved 100% accuracy on the Filesystem Operations tasks include:

- GLM-4.6
- Claude-Sonnet-4.5
- Gemini-3-Pro-Preview
- Gemini-2.5-Pro
- Gemini-2.5-Flash
- Claude-3.5-Haiku-20241022
- Llama-3.3-70B-Instruct-FP8-KV
- Qwen3-Max-Preview
- Qwen2.5-72B-Instruct
- Qwen2.5-32B-Instruct
- Claude-Haiku-4.5

Text Search and Extraction

In the third task category, Text Search and Extraction, more varied model performance begins to appear. Three models, GPT-5, GPT-5.2, and Claude-Sonnet-4.5 scored above 90% on average across the four tasks. The top 20 highest performing models can be seen in Figure 8.

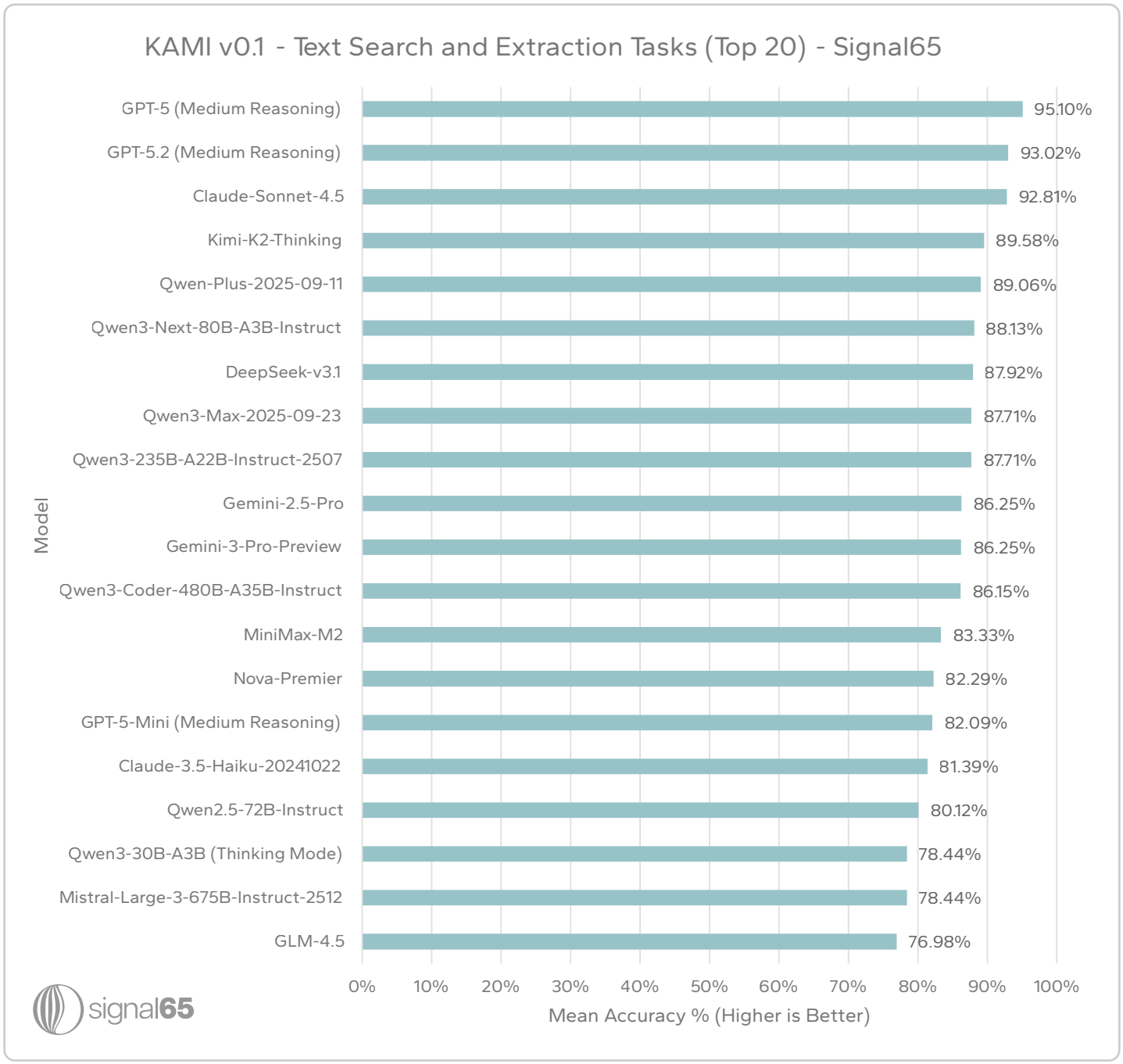


Figure 8: *Text Search and Extraction (Top 20)*

In general, models were much more successful in the first two tasks, retrieval of specific lines, compared to the second two tasks which require retrieval of specific words. The average accuracy across all models was 38.37% and 32.81% for the two word retrieval tasks, with no model achieving 100% accuracy for either task. Several models receiving lower scores achieved 90% or more on the line retrieval tasks, while receiving far lower on the word retrieval tasks, in some cases less than 1%. One of the most notable examples of this trend is GLM-4.6, which achieved the second highest overall mean accuracy across all tests, but only achieved 75.10% accuracy for text search and extraction. GLM-4.6 achieved 100% and 99.58% accuracy for the two line retrieval tasks, while achieving only 67.08% and 33.75% for the two word retrieval tasks.

CSV Processing

In the initial KAMI testing, the CSV Processing tasks were found to be amongst the most challenging. The expanded test set includes several new models achieving above 90% accuracy, led by MiniMax-M2 at 97.6%. As can be seen in Figure 9, however, accuracy declines significantly after the top 10 models.

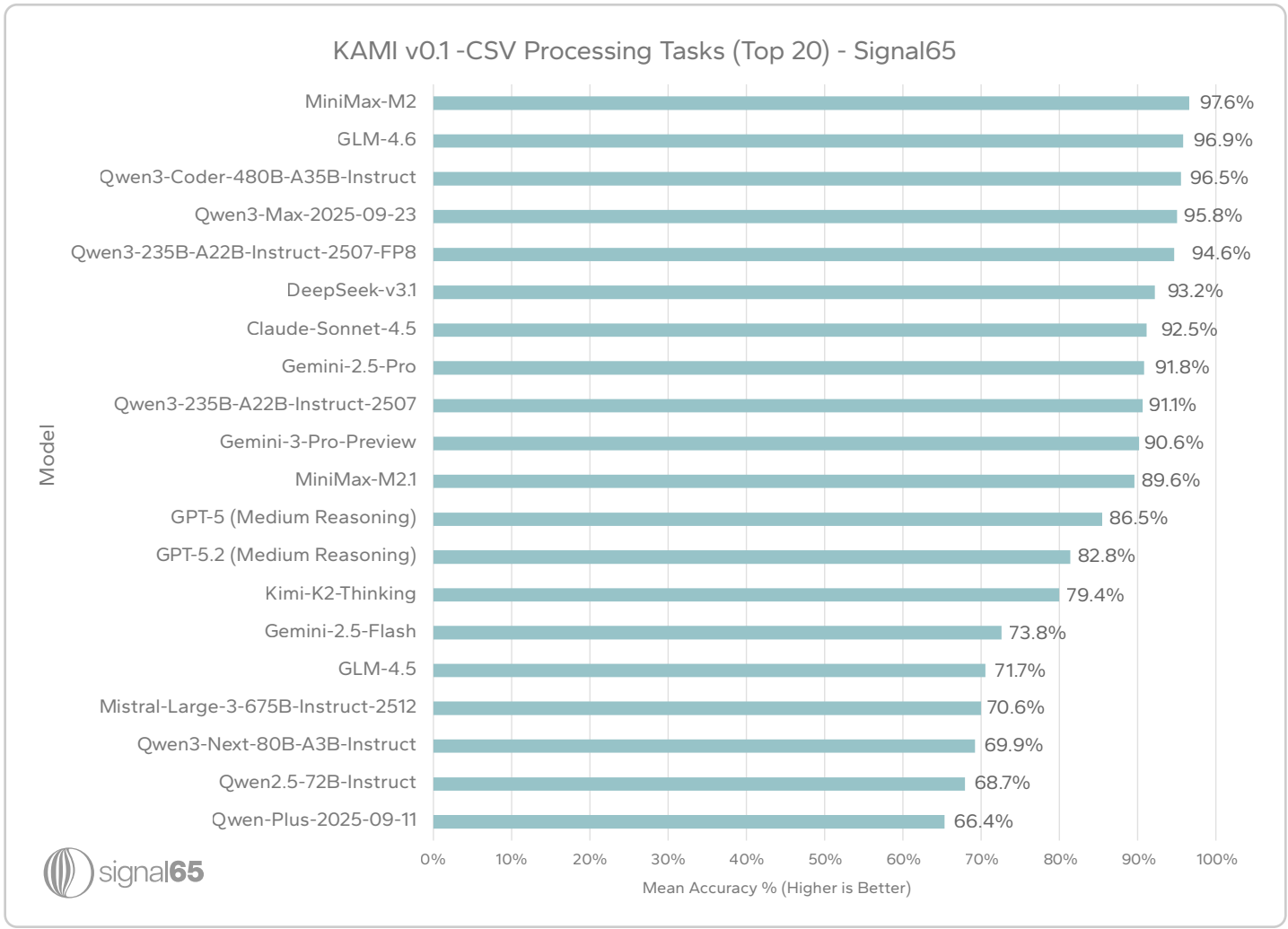


Figure 9: *CSV Processing Results (Top 20)*

For many models, this decline is attributed to the second CSV processing task. While the first question asks models to retrieve CSV data and answer a single question, the second requires them to retrieve data to answer six distinct questions. On average across all models, this task was completed with 29.17% accuracy, far lower than the 57.65% and 49.90% accuracies seen in the first and third CSV processing tasks.

Notably, the CSV Processing category was the only category in which the overall benchmark leader, GPT-5, achieved an average accuracy below 90%. While GPT-5 was highly accurate for the first and third tasks, achieving accuracies of 97.9% and 99.6%, it was challenged by the complex second task, achieving only 62.1% accuracy. This trend was seen across many other models, and can be observed even within the top 10 models in this category. As can be seen in Figure 10, the top 5 models achieve fairly consistent accuracies across all three questions, while the remaining models begin losing accuracy on the second task.

Model	CSV Task #1	CSV Task #2	CSV Task #3	Average
MiniMax-M2	100%	94.58%	98.33%	97.6%
GLM-4.6	92.92%	97.92%	100%	96.9%
Qwen3-Coder-480B-A35B-Instruct	100%	90%	99.58%	96.5%
Qwen3-Max-2025-09-23	99.58%	88.33%	99.58%	95.8%
Qwen3-235B-A22B-Instruct-2507-FP8	94.17%	94.17%	95.42%	94.6%
DeepSeek-v3.1	100%	80.83%	98.75%	93.2%
Claude-Sonnet-4.5	100%	77.5%	100%	92.5%
Gemini-2.5-Pro	98.75%	78.33%	98.33%	91.8%
Qwen3-235B-A22B-Instruct-2507	98.75%	77.08%	97.5%	91.1%
Gemini-3-Pro-Preview	100%	71.67%	100%	90.6%

Figure 10: *CSV Processing Top 5 Models*

Database Processing Tasks

In the standard Database Processing tasks, GPT-5 reclaims a strong advantage with 95% accuracy. GLM-4.6 was the only other model to achieve 90% or above.

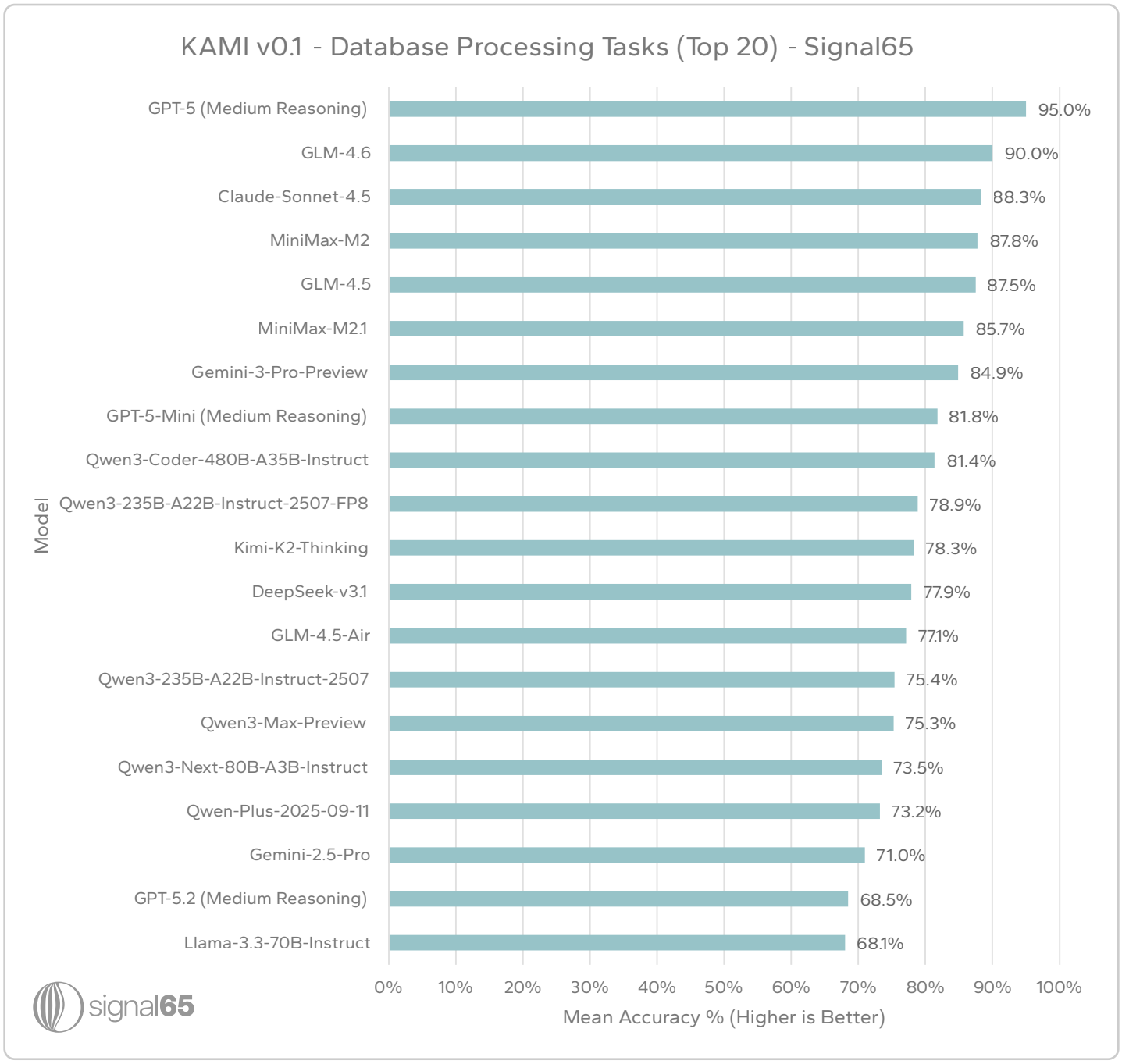


Figure 11: *Database Processing Tasks (Top 20)*

Throughout the models tested, these tasks have proven to pose several challenges. Models were often found to skip basic logical steps such as finding a table's schema before attempting to answer the questions. Some models were additionally found to return incorrect results after confusing distinct numerical columns, such as order numbers and order IDs. Models avoiding these problems demonstrate superior reasoning abilities, showing potential to handle dynamic tasks without explicit instructions.

As with the CSV Processing tasks, the second database processing task – which asks a series of 6 questions – has consistently shown to be the most challenging. While GPT-5 achieves the highest overall score due to its consistency across all three tasks, several other models – including GLM-

4.6, and Gemini-3-Pro-Preview – achieved high accuracy for tasks #1 and #3, while achieving significantly lower accuracy on task #2. GPT-5-Mini was the only other model to score above 90% on the second database task, however, it scored significantly lower on the third database task, with only 56% accuracy.

Model	DB Task #1	DB Task #2	DB Task #3	Average
GPT-5 (Medium Reasoning)	95%	90%	100%	95.0%
GLM-4.6	100%	70%	100%	90.0%
Claude-Sonnet-4.5	90%	75%	100%	88.3%
MiniMax-M2	99.17%	65%	99.17%	87.8%
GLM-4.5	100%	63.33%	99.17%	87.5%
MiniMax-M2.1	87.5%	69.58%	100%	85.7%
Gemini-3-Pro-Preview	100%	54.58%	100%	84.9%
GPT-5-Mini (Medium Reasoning)	96.67%	92.5%	56.25%	81.8%
Qwen3-Coder-480B-A35B-Instruct	97.92%	47.08%	99.17%	81.4%
Qwen3-235B-A22B-Instruct-2507-FP8	70.42%	75.42%	90.83%	78.9%

Figure 12: *Database Processing Detailed Results*

Database Processing Tasks (Guided)

The guided database tasks repeat the first two database tasks with the inclusion of hints to avoid common mistakes, such as explicit instruction to first examine the schema. The results of these tests showcase that some models can achieve notably improved performance when given more detailed prompts.

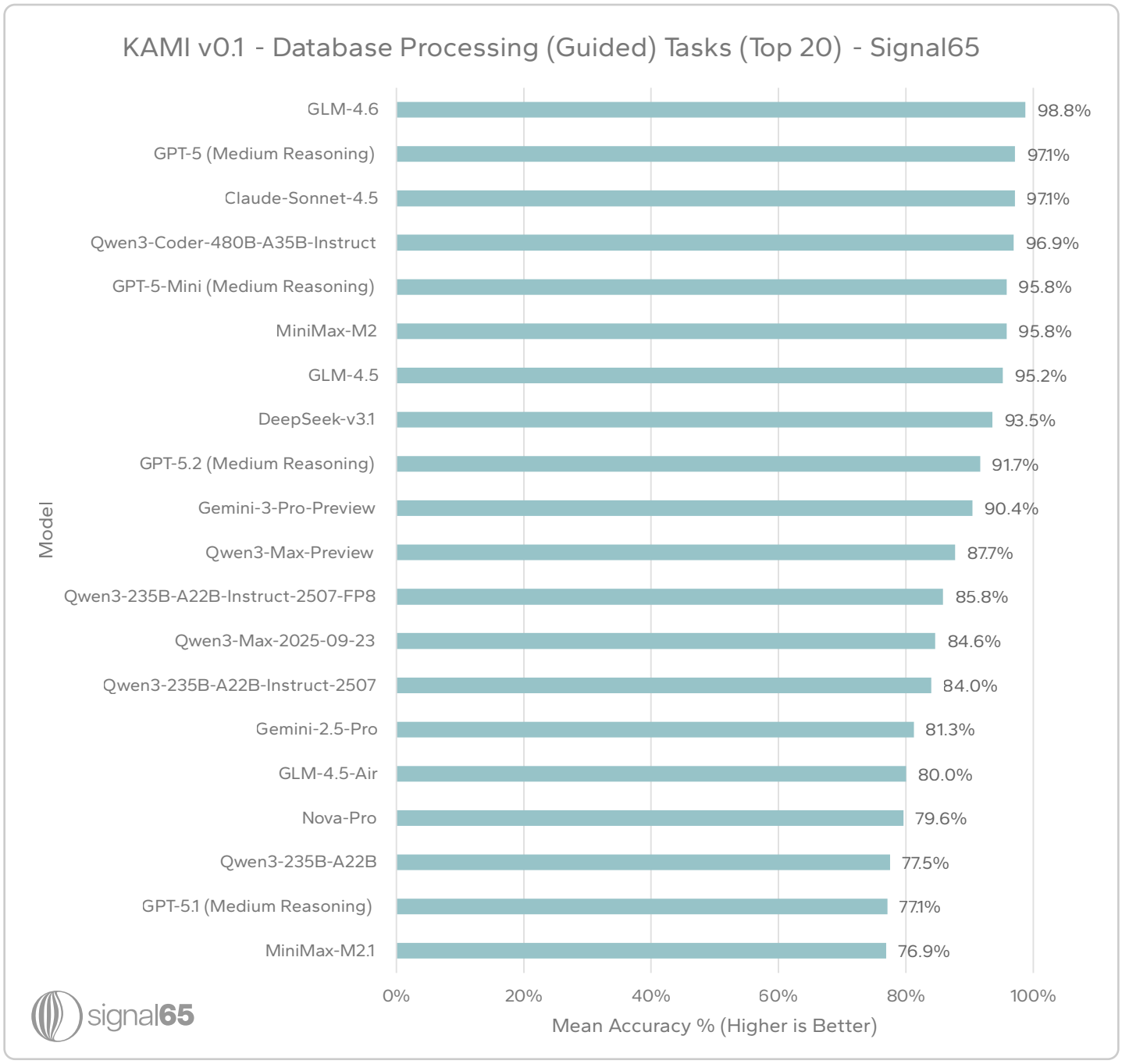


Figure 13: *Guided Database Processing Tasks (Top 20)*

While the two leaders of the standard database tasks remain for the guided tasks, GLM-4.6 notably surpasses GPT-5 when given hints, with an accuracy of 98.8%. Several other models achieved significant improvements, with ten models achieving scores of 90% or higher.

Instruction Following

The Instruction Following tasks were found to be easily achievable by most high performing models, with 19 distinct models achieving 100% accuracy, and an additional 7 achieving 99% or higher. These tasks measure the model's ability to correctly follow instructions for outputting results in various formats. For enterprise agentic applications, this level of instruction following should be considered a baseline requirement for model selection to ensure data is output correctly.

The following models all achieved 100% accuracy on the three Instruction Following tasks:

- GPT-5 (Medium Reasoning)
- GLM-4.6
- DeepSeek-v3.1
- Qwen3-235B-A22B-Instruct-2507
- Qwen3-235B-A22B-Instruct-2507-FP8
- Llama-3.3-70B-Instruct
- Llama-3.3-70B-Instruct-FP8-KV
- Qwen3-Max-Preview
- Llama-4-Maverick-17B-128E-Instruct-FP8
- Qwen3-30B-A3B (Thinking Mode)
- Qwen3-30B-A3B-Instruct-2507
- Qwen3-14B (Thinking Mode)
- Qwen-Flash-2025-07-28
- Qwen3-235B-A22B
- Qwen2.5-14B-Instruct
- Llama-4-Scout-17B-16E-Instruct
- Qwen3-4B-Instruct-2507
- Qwen3-14B
- Qwen3-30B-A3B

While many models were found to be generally successful in achieving the Instruction Following tasks, a few otherwise high performing models stand out with uncharacteristically low performance. This was often attributed to the second task, in which models were tasked with providing their response in JSON format. Examples of this behavior can be seen in Figure 14.

Model	Instruction Following Task #1	Instruction Following Task #2	Instruction Following Task #3	Average
MiniMax-M2	98.33%	30%	100%	76.1%
Gemini-3-Pro-Preview	100%	17.92%	100%	72.6%
Gemini-2.5-Pro	100%	0%	100%	66.7%
Kimi-K2-Thinking	93.89%	38.33%	58.89%	63.7%

Figure 14: *Instruction Following Tasks Detailed Results*

AWS Bedrock

As previously mentioned, some models were tested both on hardware within the Signal65 AI lab and again on AWS Bedrock. Interestingly, some models achieved higher accuracy when run on AWS Bedrock.

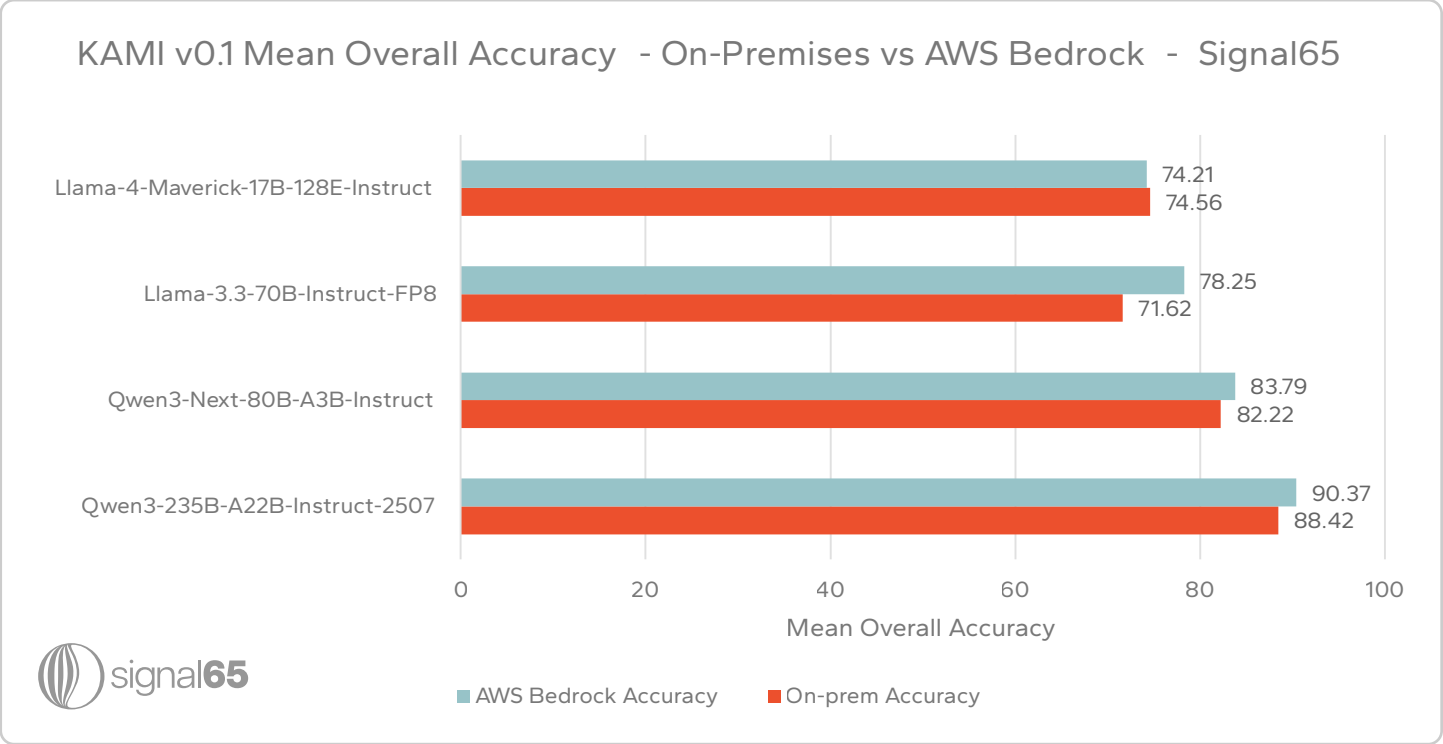


Figure 15: *On-Premises vs AWS Bedrock Accuracy*

Due to the randomized approach of the KAMI Benchmark, some variance between runs is expected, which is mitigated by repeated runs of the benchmark. In some cases, however, such as Llama-3-3-70B-Instruct-FP8-KV and Qwen3-235B-A22B-Instruct-2407, the difference between accuracies was found to be statistically significant, with AWS Bedrock enabling higher overall accuracy. The most noticeable difference is found in Llama 3.3-50B-Instruct-FP8, which improves from 71.62% to 78.25% when run on AWS Bedrock.

These improvements are interesting, and may be attributed to several variables, however without specific knowledge of AWS Bedrock configurations, the exact causes remain unclear. Possible explanations include how models are hosted, the hardware utilized, or any specific prompts and guardrails implemented by AWS.

While these models do show some interesting improvements, it should be noted that this is a small sample size and it cannot be concluded that AWS Bedrock provides higher accuracies in all circumstances. As can be seen in Figure 15, Llama-4-Maverick-17B-128E-Instruct demonstrates an example of a model achieving nearly identical accuracy both on-premises and on AWS Bedrock, with the model achieving slightly higher accuracy on-premises.

These results do show, however, that implementation details can make a notable difference in accuracy, and present an additional area for further experimentation.

Proprietary Models

While previous KAMI testing primarily focused on open source models run in the Signal65 AI Lab, testing in Q1 2026 notably expanded the test set with several popular proprietary models. Proprietary models present an interesting dynamic between accessibility and cost. Proprietary models from OpenAI, Anthropic, and Google are often heavily leveraged by enterprise organizations due to their ease of access over an API. These proprietary models enable organizations to avoid the complexity of managing infrastructure and model deployment; however, they come with ongoing API costs – often charged per token.

Open source models on the other hand, can be deployed and run without ongoing API fees, but require the upfront cost and complexity of deployment. Incorporating proprietary models into the KAMI test set, provides organizations a way to weigh these considerations against agentic performance. Figure 16 shows the overall results for all proprietary models tested.

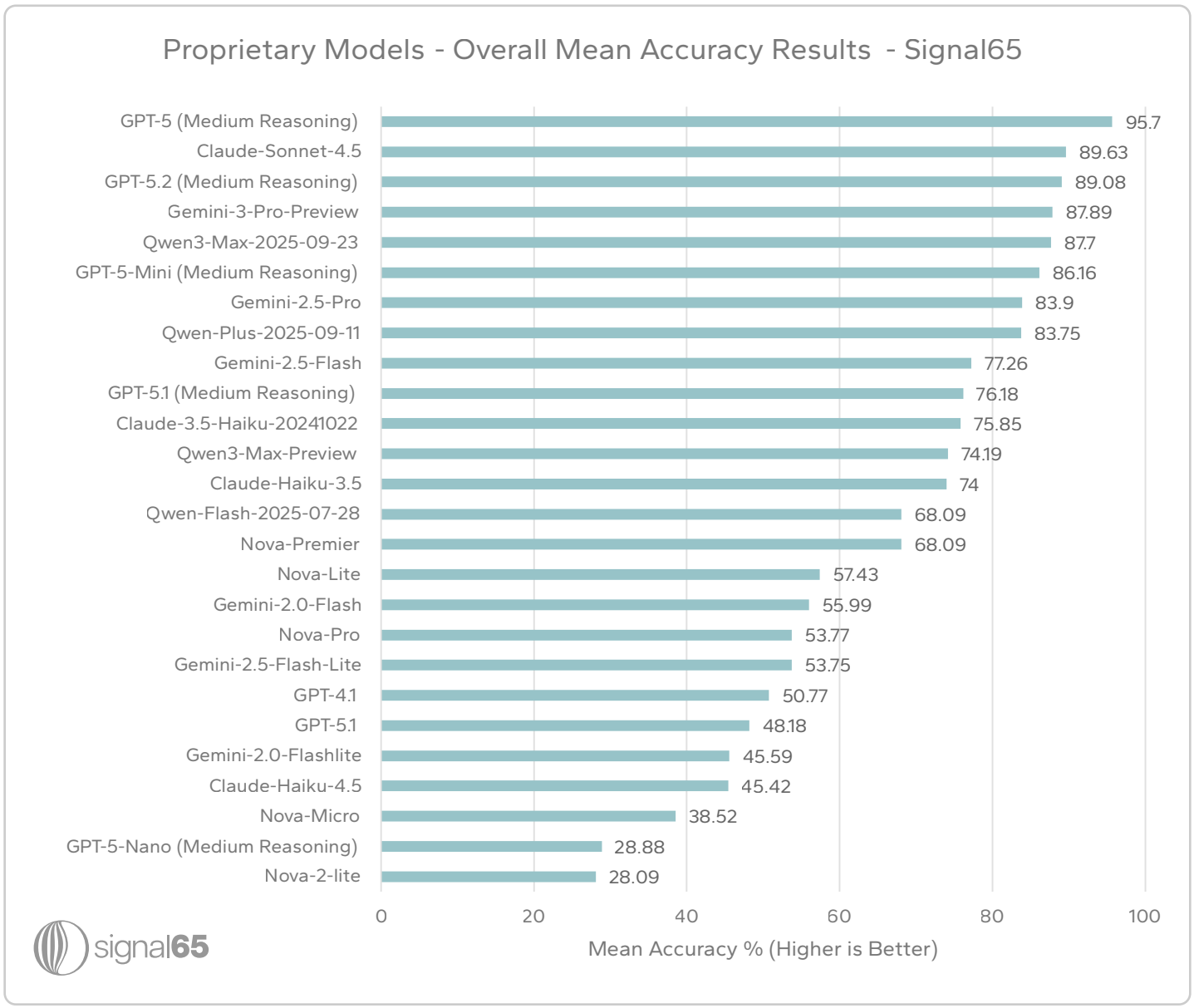


Figure 16: *Proprietary Models Overall Mean Accuracy*

Most notable amongst these results is GPT-5, which not only leads all proprietary models tested, but achieved the highest overall score of all models at 95.7% accuracy. The remaining top 5 proprietary models – Claude-Sonnet-4.5, Gemini-3-Pro-Preview, Qwen3-Max-2025-09-23, and GPT-5-Mini – additionally achieved relatively high scores ranging from 86.16% to 89.63%. When compared to leading open source models, however, there does not appear to be a notable gap between proprietary and open source models. Figure 17 shows the top 10 highest performing open source models.

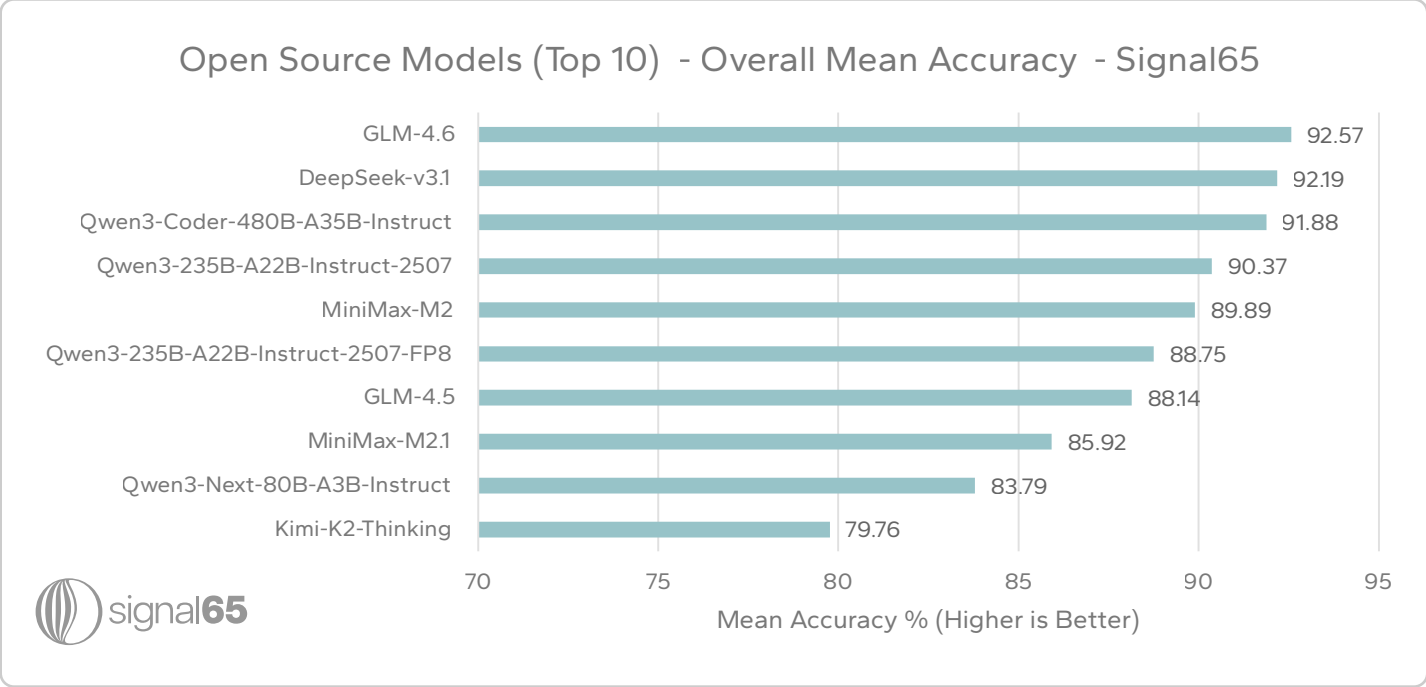


Figure 17: *Top 10 Open Source Models*

Interestingly, the top five open source models outperform all proprietary models tested, with the exception of GPT-5. These results demonstrate that for agentic applications, open source models are capable of performing as well, or better than leading proprietary models.

The Qwen models in particular provide an interesting view into the state of both proprietary and open source models, as the only provider with both types of models tested. The highest performing proprietary Qwen model, Qwen3-Max-2025-09-23, scores well with 87.7% mean overall accuracy, making it the twelfth highest performing model tested. However, multiple open source Qwen models achieved even higher scores, including Qwen3-Coder-480B-A35B-Instruct, and Qwen3-235B-A22B-Instruct-2507 (both FP8 and full-weight versions). This shows that even within a prominent model family such as Qwen, open source models can compete or even surpass proprietary options.

This competitive performance presents significantly more options, as well as potential economic savings, to enterprise organizations who primarily rely on proprietary models. Agentic workloads are typically iterative, requiring models to alternate between reasoning and calling tools, until a given task is completed. Additionally, many enterprise tasks are highly repetitive, requiring agents to complete specific jobs repeatedly. Due to this, agentic AI can result in an extremely high number of inferences and total tokens. When considering proprietary models that often charge per token – this can amount to prohibitively high costs.

High performing open source models, such as GLM-4.6, can provide an alternative approach, enabling organizations to achieve high agentic accuracy without ongoing per-token costs. It should be noted, however, that many of the highest performing open source models are very large – requiring significant infrastructure, as well as energy costs, to run on-premises. Amongst the top 10 open source models, Qwen3-Next-80B-A3B-Instruct presents significant value as an 80 billion parameter model. While this is still a large model, it is relatively small compared to the other top performing open source models, which range from Qwen3-235B-A22B-Instruct-2507-FP8 at 235 billion parameters to Kimi-K2-Thinking at 1 trillion parameters.

To fully understand the economic tradeoffs of either approach, organizations should conduct a TCO analysis for their specific agentic AI requirements, however, the KAMI v0.1 results show leading models in both categories can achieve highly accurate results. A brief overview of approximate costs for the top three proprietary models tested is shown in Figure 18 and Figure 19 utilizing the average input and output tokens per conversation observed during testing. This can be used to approximate the average cost per token and extrapolated to calculate the approximate cost for a single run of the KAMI v0.1 Benchmark, with 570 total conversations.

Model	Input Token Cost	Output Token Cost	Average Input Tokens per Conversation	Approximate Input Cost per Conversation
GPT-5	$1.25/M Tokens	$10/M Tokens	21,436.88	$0.026796103
Claude-Sonnet-4.5	$3/M Tokens	$15/M Tokens	24,554.18	$0.07366254
GPT-5.2	$1.75/M Tokens	$14/M Tokens	42,536.44	$0.074438774

Model	Average Output Tokens per Conversation	Approximate Output Cost per Conversation	Approximate Cost per Conversation	Approximate Cost per Test Run (570 Conversation)
GPT-5	449.34	$0.004493	$0.031289491	$17.83500966
Claude-Sonnet-4.5	688.61	$0.010329094	$0.083991634	$47.87523124
GPT-5.2	1955.48	$0.027377	$0.101815494	$58.03483179

Figures 18 & 19: *Approximate API Costs*

Signal65 Comment *– These calculations represent rough approximations for demonstrative purposes only. Notably, these calculations simplify model API cost structures, excluding more complex pricing such as cached token pricing, or various priority tiers. Costs may additionally vary significantly between specific enterprise use cases. These calculations are only intended as a representation of how significant token usage during agentic workloads can impact API costs, and should not be used to for financial planning of agentic workloads.*

Model Families

The rapid pace of AI development has led to models of various versions and sizes within single model families. In general, it is assumed that larger and newer models should outperform older and smaller models. When testing open source models, it was seen that this was not always the case. Previous KAMI testing found examples of older model variations outperforming newer versions across both Qwen and Llama model families. Similar insights can be gained by examining distinct models tested within proprietary model families.

GPT

The GPT family shows interesting results across various versions, sizes, and reasoning capabilities. GPT-5, as previously noted, stands out as the top performing model tested. The 95.7% accuracy shows a notable advancement over the previous generation GPT-4.1 at only 50.77% accuracy. Curiously, it also significantly outperforms the newer GPT-5.1 and GPT-5.2 models. The smaller GPT-5-Mini variation, additionally stands out with 86.16% performance, which also outperforms the GPT-5.1 models tested.

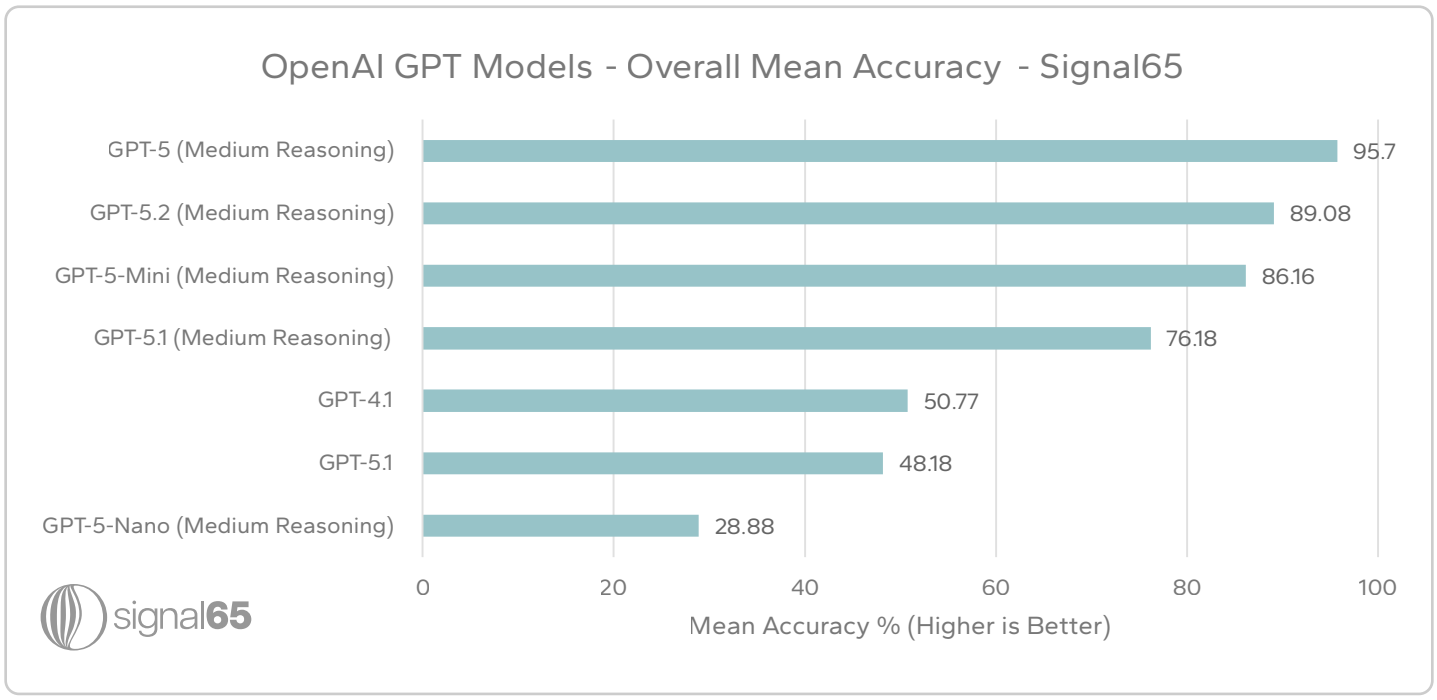


Figure 20: *GPT Model Family Overall Mean Accuracy*

The base GPT-5.1 model, run with its default settings (non-reasoning), achieved below 50% accuracy, a notable regression from GPT-5, and additionally underperforming GPT-4.1. A key distinction between GPT-5.1 and GPT-5 is how reasoning is configured. As a default, GPT-5 includes reasoning, while GPT-5.1 does not. GPT-5.1 was additionally run with medium reasoning abilities enabled, and while this significantly improved performance from 48.18% to 76.18%, it is still well below the previous GPT-5 and GPT-5-Mini models.

Gemini

While at its max, Gemini models do not achieve the accuracy of GPT-5, the model family shows consistent advancement.

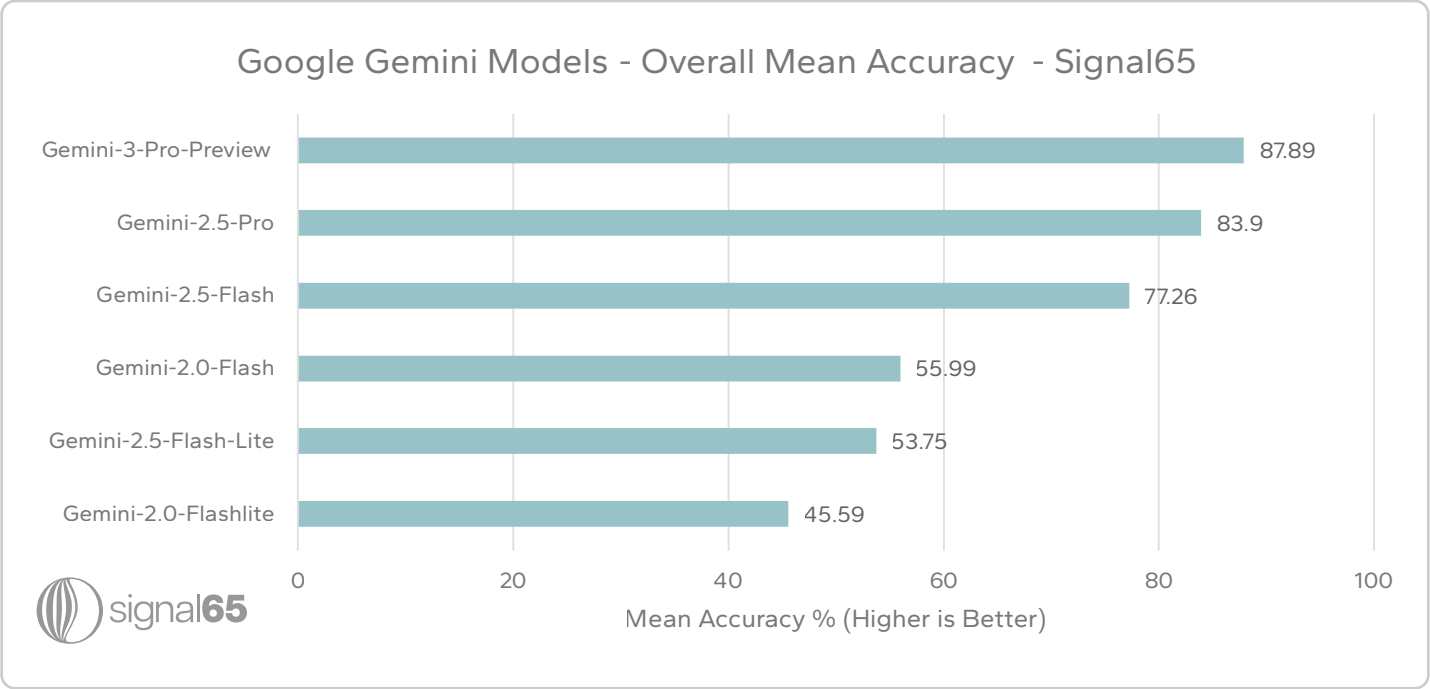


Figure 21: *Gemini Model Family Overall Mean Accuracy*

Gemini-3-Pro-Preveiw leads the model family with 87.89%, as would be logically expected as the newest iteration. Similarly, the Gemini-2.5 Pro and Flash models show notable improvement on the previous generation Gemini-2 models. Gemini-2.5-Flash-Lite was found to slightly underperform the previous generation Gemini-2.0-Flash, however this is likely attributed to its smaller size. Gemini-2.5-Flashlight was still seen to achieve a notable improvement over Gemini-2.0-Flashlight.

Claude

Anthropic's Claude models are led by Claude-Sonnet-4.5, which scored 89.63%, making it one the five most accurate model tested, and one of the only three proprietary models included in the top ten. Logically, Claude-Sonnet outperforming Claude-Haiku models makes sense, as the Sonnet models are intended to be larger models with greater reasoning capabilities, attributes that have consistently shown to be beneficial in KAMI agentic scenarios.

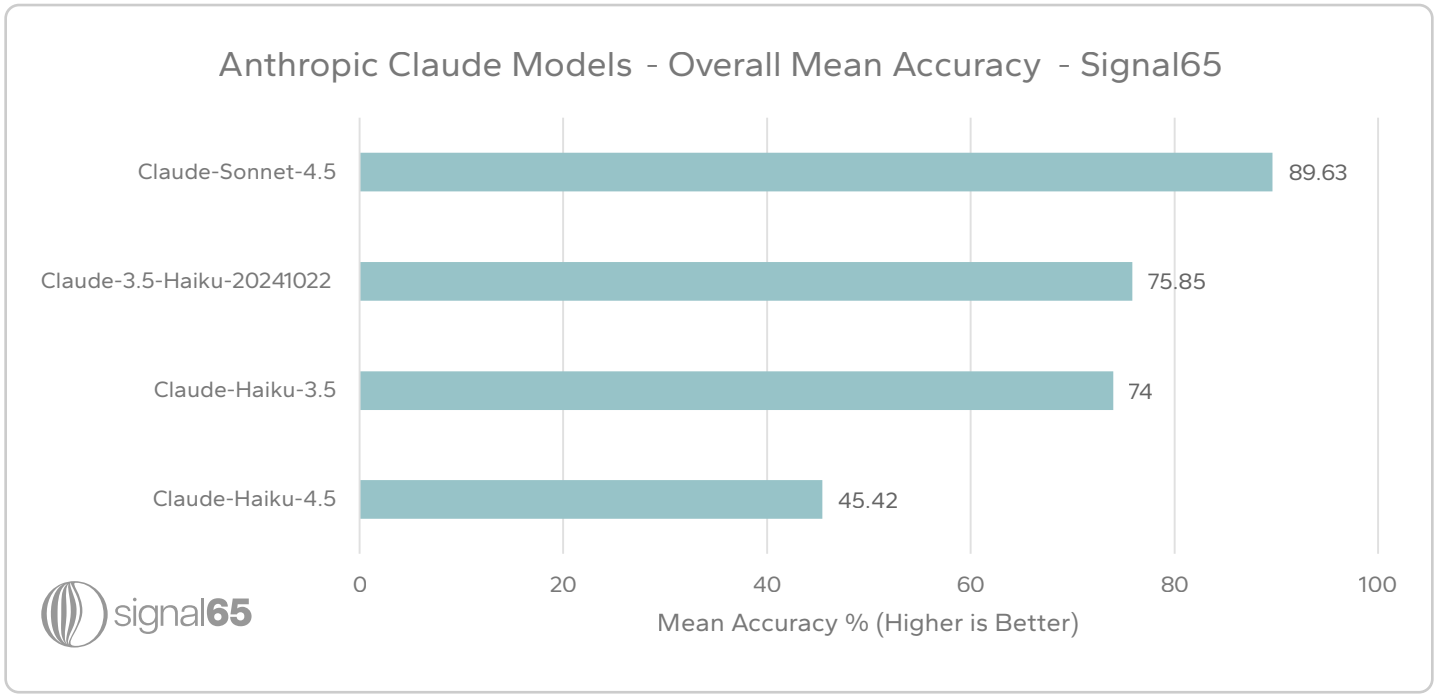


Figure 22: *Claude Model Family Overall Mean Accuracy*

Examining the Claude-Haiku models tested, however, again shows interesting progression within a model family. Claude-3.5-Haiku-20241022 outperforms the standard Claudde-Haiku-3.5, however only slightly, showing little noticeable improvement. More notable, however, is that Claude-Haiku-4.5 significantly underperforms both Claude-3.5 iterations, with only 45.42% overall mean accuracy.

Nova

In general, the Amazon Nova model family was found to achieve relatively low accuracy compared to competitive models, both proprietary and open source. The top performing Nova model was Nova-Premier, which achieved 68.09% accuracy.

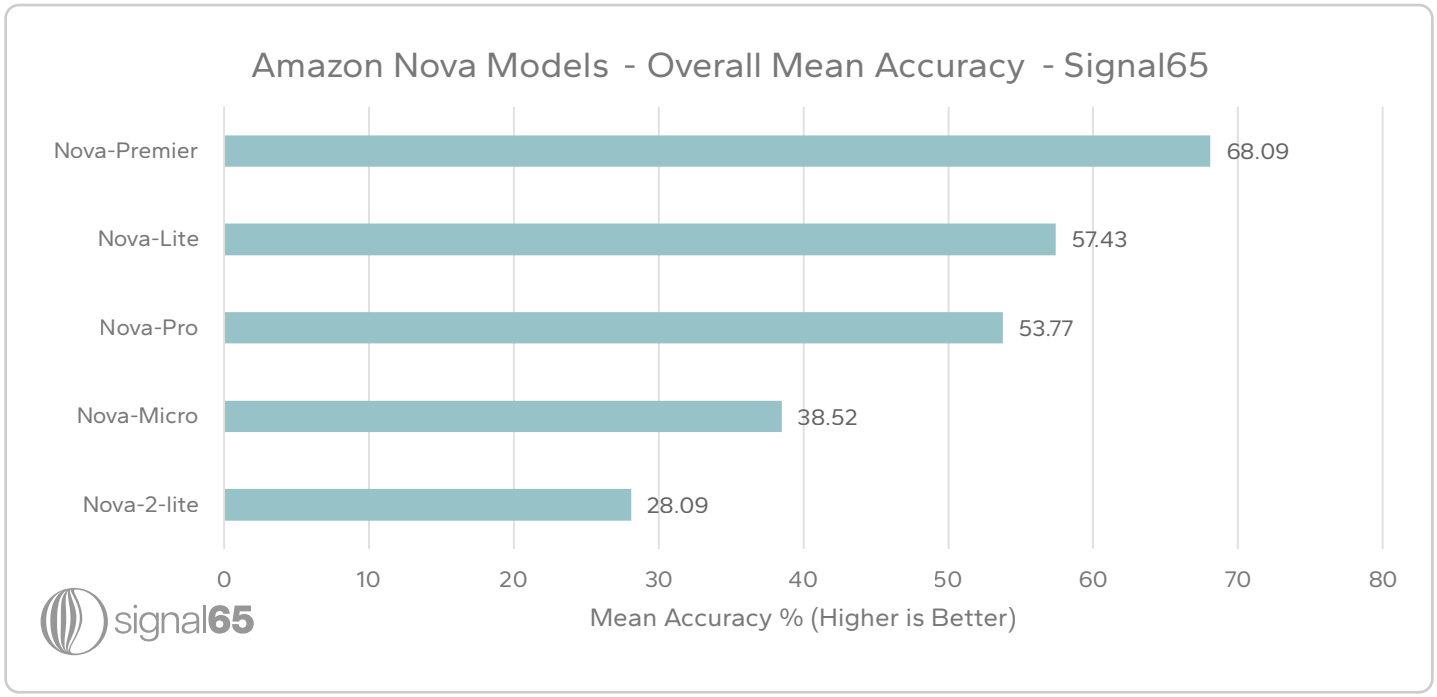


Figure 23: *Nova Model Family Overall Mean Accuracy*

The remaining Nova models show inconsistent performance variations between both model size and model version. Nova-Lite was found to outperform Nova-Pro, while Nova-2-Lite trails all other Nova models at only 28.09% accuracy, the 4th lowest score of all models tested.

Final Thoughts: Uncovering Top Agentic Models with KAMI

The KAMI benchmark provides a realistic look into the agentic capabilities of leading AI models, testing real world agentic tasks not seen in other popular AI benchmarks. This second iteration of testing with the KAMI v0.1 Benchmark builds upon previous findings and notably includes several prominent proprietary model families.

The inclusion of both open source and proprietary models is crucial to create a comprehensive understanding of the AI landscape. This testing demonstrates that certain proprietary models, such as GPT-5 stand out amongst the AI landscape, while additionally challenging the idea that all proprietary models are inherently superior. Several open source models, led by GLM-4.6, were found to outperform leading proprietary models, indicating that there is not a significant gap between open source and proprietary options. These results showcase that enterprises have a wide variety of options, and open discussion around the economic practicality of both approaches.

Key Highlights

- GPT-5 is the top agentic AI performer at **95.7% mean accuracy score**
- **GLM-4.6 leads all open models** with 92.57% mean accuracy
- Open models achieve **7 of the top 10** highest accuracies for agentic workloads

These results additionally highlight inconsistencies in ongoing model development. While model developers are seemingly in a race to constantly produce newer, better models, results from the KAMI v0.1 Benchmark indicate that not all new releases make tangible improvements when considering agentic workloads. As seen in past open source results, as well as within proprietary model families, some newer model iterations actually perform worse during agentic tasks than their previous versions. This may be indicative of model development being guided by flawed AI benchmarks, which enable memorization to dictate perceived performance. Signal65 believes that efforts to create new benchmarks focused on agentic workloads, such as KAMI, will help drive the industry forward by uncovering how models perform during real world agentic use cases.

Signal65 and Kamiwaza will continue to iterate on the KAMI Benchmark, testing additional models, as well as enhancing the test suite to provide insights into the agentic AI landscape.

Appendix

Overall Mean Accuracy

Model	Mean Accuracy
GPT-5 (Medium Reasoning)	95.7%
GLM-4.6	92.57%
DeepSeek-v3.1	92.19%
Qwen3-Coder-480B-A35B-Instruct	91.88%
Qwen3-235B-A22B-Instruct-2507	90.37%
MiniMax-M2	89.89%
Claude-Sonnet-4.5	89.63%
GPT-5.2 (Medium Reasoning)	89.08%
Qwen3-235B-A22B-Instruct-2507-FP8	88.75%
GLM-4.5	88.14%
Gemini-3-Pro-Preview	87.89%
Qwen3-Max-2025-09-23	87.7%
GPT-5-Mini (Medium Reasoning)	86.16%
MiniMax-M2.1	85.92%
Gemini-2.5-Pro	83.9%
Qwen3-Next-80B-A3B-Instruct	83.79%
Qwen-Plus-2025-09-11	83.75%
Kimi-K2-Thinking	79.76%
Llama-3.3-70B-Instruct	78.25%
Gemini-2.5-Flash	77.26%
GPT-5.1 (Medium Reasoning)	76.18%
Claude-3.5-Haiku-20241022	75.85%
GLM-4.5-Air	75.33%

Mistral-Large-3-675B-Instruct-2512	74.98%
Llama-4-Maverick-17B-128E-Instruct	74.56%
Llama-3.3-70B-Instruct-FP8-KV	74.54%
Qwen3-Max-Preview	74.19%
Claude-Haiku-3.5	74%
Llama-3.1-70B-Instruct	73.44%
Llama-4-Maverick-17B-128E-Instruct-FP8	73.14%
Qwen3-30B-A3B (Thinking Mode)	72.74%
Qwen2.5-72B-Instruct	71.08%
Qwen3-30B-A3B-Instruct-2507	69.65%
Qwen3-14B (Thinking Mode)	69.1%
Qwen-Flash-2025-07-28	68.09%
Nova-Premier	68.09%
Qwen3-235B-A22B	67.65%
Qwen3-32B (Thinking Mode)	67.61%
Qwen2.5-14B-Instruct	66.56%
Llama-4-Scout-17B-16E-Instruct	64.06%
Qwen3-32B-FP8	63.71%
Qwen3-8B (Thinking Mode)	62.54%
Qwen3-32B	61.56%
Qwen3-14B-FP8	60%
Qwen3-4B-Instruct-2507	60%
DeepSeek-v3	59.36%
Mistral-Large-Instruct-2411	58.9%
Qwen3-14B	58.75%
Granite-4.0-H-Small	58.51%
Qwen3-30B-A3B	58.11%
Nova-Lite	57.43%

Gemini-2.0-Flash	55.99%
Qwen2.5-32B-Instruct	55.9%
Phi-4	54.81%
Nova-Pro	53.77%
Gemini-2.5-Flash-Lite	53.75%
GPT-4.1	50.77%
Qwen3-4B (Thinking Mode)	50.53%
Qwen3-8B	49.05%
GPT-5.1	48.18%
Gemini-2.0-Flashlite	45.59%
Claude-Haiku-4.5	45.42%
Qwen2.5-7B-Instruct	41.56%
Nova-Micro	38.52%
Qwen3-4B	37.78%
GPT-5-Nano (Medium Reasoning)	28.88%
Nova-2-lite	28.09%
Granite-4.0-H-Tiny	27.26%
Granite-4.0-H-Micro	17.06%
Llama-3.1-8B-Instruct	10.5%

Basic Reasoning Results

Model	Q101	Q102	Average
GPT-5 (Medium Reasoning)	100.00%	100.00%	100.00%
GLM-4.6	100.00%	100.00%	100.00%
DeepSeek-v3.1	100.00%	100.00%	100.00%
Qwen3-235B-A22B-Instruct-2507	100.00%	100.00%	100.00%
MiniMax-M2	100.00%	100.00%	100.00%
GPT-5.2 (Medium Reasoning)	100.00%	100.00%	100.00%
Qwen3-235B-A22B-Instruct-2507-FP8	100.00%	100.00%	100.00%
GLM-4.5	100.00%	100.00%	100.00%
Gemini-3-Pro-Preview	100.00%	100.00%	100.00%
Qwen3-Max-2025-09-23	100.00%	100.00%	100.00%
GPT-5-Mini (Medium Reasoning)	100.00%	100.00%	100.00%
MiniMax-M2.1	100.00%	100.00%	100.00%
Qwen3-Next-80B-A3B-Instruct	100.00%	100.00%	100.00%
Qwen-Plus-2025-09-11	100.00%	100.00%	100.00%
Llama-3.3-70B-Instruct	100.00%	100.00%	100.00%
Gemini-2.5-Flash	100.00%	100.00%	100.00%
GPT-5.1 (Medium Reasoning)	100.00%	100.00%	100.00%
Claude-3.5-Haiku-20241022	100.00%	100.00%	100.00%
Llama-4-Maverick-17B-128E-Instruct	100.00%	100.00%	100.00%
Llama-3.3-70B-Instruct-FP8-KV	100.00%	100.00%	100.00%
Qwen3-Max-Preview	100.00%	100.00%	100.00%
Claude-Haiku-3.5	100.00%	100.00%	100.00%
Llama-3.1-70B-Instruct	100.00%	100.00%	100.00%
Llama-4-Maverick-17B-128E-Instruct-FP8	100.00%	100.00%	100.00%
Qwen3-30B-A3B (Thinking Mode)	100.00%	100.00%	100.00%

Qwen2.5-72B-Instruct	100.00%	100.00%	100.00%
Qwen3-30B-A3B-Instruct-2507	100.00%	100.00%	100.00%
Qwen3-14B (Thinking Mode)	100.00%	100.00%	100.00%
Qwen-Flash-2025-07-28	100.00%	100.00%	100.00%
Qwen3-235B-A22B	100.00%	100.00%	100.00%
Qwen2.5-14B-Instruct	100.00%	100.00%	100.00%
Llama-4-Scout-17B-16E-Instruct	100.00%	100.00%	100.00%
Qwen3-32B-FP8	100.00%	100.00%	100.00%
Qwen3-8B (Thinking Mode)	100.00%	100.00%	100.00%
Qwen3-32B	100.00%	100.00%	100.00%
Qwen3-14B-FP8	100.00%	100.00%	100.00%
Qwen3-4B-Instruct-2507	100.00%	100.00%	100.00%
DeepSeek-v3	100.00%	100.00%	100.00%
Qwen3-14B	100.00%	100.00%	100.00%
Qwen3-30B-A3B	100.00%	100.00%	100.00%
Qwen2.5-32B-Instruct	100.00%	100.00%	100.00%
Gemini-2.5-Flash-Lite	100.00%	100.00%	100.00%
GPT-4.1	100.00%	100.00%	100.00%
Qwen3-8B	100.00%	100.00%	100.00%
GPT-5.1	100.00%	100.00%	100.00%
Claude-Haiku-4.5	100.00%	100.00%	100.00%
Qwen2.5-7B-Instruct	100.00%	100.00%	100.00%
Claude-Sonnet-4.5	99.58%	100.00%	99.79%
Mistral-Large-Instruct-2411	100.00%	99.58%	99.79%
GPT-5-Nano (Medium Reasoning)	99.58%	100.00%	99.79%
Nova-2-lite	100.00%	99.58%	99.79%
GLM-4.5-Air	99.33%	100.00%	99.67%
Qwen3-Coder-480B-A35B-Instruct	100.00%	99.17%	99.59%

Gemini-2.0-Flash	99.17%	100.00%	99.59%
Qwen3-4B	100.00%	98.89%	99.45%
Qwen3-4B (Thinking Mode)	98.75%	99.58%	99.17%
Gemini-2.5-Pro	98.33%	100.00%	99.17%
Kimi-K2-Thinking	98.33%	100.00%	99.17%
Qwen3-32B (Thinking Mode)	97.92%	100.00%	98.96%
Gemini-2.0-Flashlite	100.00%	97.92%	98.96%
Granite-4.0-H-Small	97.08%	99.17%	98.13%
Mistral-Large-3-675B-Instruct-2512	98.75%	96.67%	97.71%
Granite-4.0-H-Micro	98.33%	91.25%	94.79%
Phi-4	98.57%	87.62%	93.10%
Nova-Premier	82.08%	99.17%	90.63%
Granite-4.0-H-Tiny	83.75%	80.00%	81.88%
Nova-Lite	81.67%	54.17%	67.92%
Nova-Pro	15.83%	55.83%	35.83%
Llama-3.1-8B-Instruct	30.56%	24.44%	27.50%
Nova-Micro	44.76%	8.57%	26.67%

Filesystem Task Results

Model	Q201	Q202	Average
GLM-4.6	100.00%	100.00%	100.0%
Claude-Sonnet-4.5	100.00%	100.00%	100.0%
Gemini-3-Pro-Preview	100.00%	100.00%	100.0%
Gemini-2.5-Pro	100.00%	100.00%	100.0%
Gemini-2.5-Flash	100.00%	100.00%	100.0%
Claude-3.5-Haiku-20241022	100.00%	100.00%	100.0%
Llama-3.3-70B-Instruct-FP8-KV	100.00%	100.00%	100.0%
Qwen3-Max-Preview	100.00%	100.00%	100.0%
Qwen2.5-72B-Instruct	100.00%	100.00%	100.0%
Qwen2.5-32B-Instruct	100.00%	100.00%	100.0%
Claude-Haiku-4.5	100.00%	100.00%	100.0%
DeepSeek-v3.1	100.00%	99.58%	99.8%
Qwen3-235B-A22B-Instruct-2507-FP8	100.00%	99.58%	99.8%
Qwen3-Max-2025-09-23	100.00%	99.58%	99.8%
Llama-3.1-70B-Instruct	99.58%	100.00%	99.8%
Claude-Haiku-3.5	99.33%	100.00%	99.7%
GPT-5 (Medium Reasoning)	100.00%	99.17%	99.6%
GLM-4.5	100.00%	99.17%	99.6%
Qwen3-32B	100.00%	99.17%	99.6%
Nova-Pro	100.00%	99.17%	99.6%
Llama-3.3-70B-Instruct	99.58%	99.58%	99.6%
Qwen3-Coder-480B-A35B-Instruct	100.00%	98.75%	99.4%
Qwen3-235B-A22B-Instruct-2507	100.00%	98.75%	99.4%
DeepSeek-v3	100.00%	98.75%	99.4%
Nova-Premier	98.75%	99.58%	99.2%

MiniMax-M2	100.00%	98.33%	99.2%
Qwen3-Next-80B-A3B-Instruct	100.00%	98.33%	99.2%
Qwen-Plus-2025-09-11	100.00%	98.33%	99.2%
Qwen3-32B-FP8	100.00%	98.33%	99.2%
GPT-5.2 (Medium Reasoning)	99.17%	98.75%	99.0%
Qwen3-30B-A3B	99.17%	98.75%	99.0%
Llama-4-Maverick-17B-128E-Instruct	100.00%	97.50%	98.8%
Llama-4-Maverick-17B-128E-Instruct-FP8	100.00%	97.50%	98.8%
GLM-4.5-Air	99.33%	98.00%	98.7%
GPT-5-Mini (Medium Reasoning)	100.00%	97.08%	98.5%
Qwen3-14B-FP8	100.00%	96.67%	98.3%
MiniMax-M2.1	99.58%	96.67%	98.1%
Qwen3-14B	100.00%	95.83%	97.9%
Granite-4.0-H-Small	98.75%	95.83%	97.3%
Qwen3-30B-A3B-Instruct-2507	100.00%	94.17%	97.1%
Qwen-Flash-2025-07-28	99.58%	92.50%	96.0%
Mistral-Large-3-675B-Instruct-2512	100.00%	90.42%	95.2%
Qwen3-235B-A22B	93.33%	95.00%	94.2%
GPT-5.1 (Medium Reasoning)	95.00%	91.67%	93.3%
Qwen2.5-14B-Instruct	99.58%	86.25%	92.9%
Qwen3-32B (Thinking Mode)	92.92%	88.75%	90.8%
Qwen3-8B	98.67%	81.33%	90.0%
Qwen3-30B-A3B (Thinking Mode)	97.92%	81.67%	89.8%
Mistral-Large-Instruct-2411	87.08%	92.08%	89.6%
Qwen3-4B-Instruct-2507	100.00%	72.50%	86.3%
Qwen3-4B	97.78%	73.33%	85.6%
GPT-5.1	91.67%	78.75%	85.2%
Nova-Lite	100.00%	69.58%	84.8%

Qwen3-8B (Thinking Mode)	97.92%	70.42%	84.2%
Phi-4	80.00%	87.14%	83.6%
Qwen3-14B (Thinking Mode)	99.17%	60.00%	79.6%
Gemini-2.5-Flash-Lite	58.75%	100.00%	79.4%
Gemini-2.0-Flash	68.75%	86.67%	77.7%
Kimi-K2-Thinking	96.67%	57.78%	77.2%
Llama-4-Scout-17B-16E-Instruct	64.17%	69.17%	66.7%
Qwen3-4B (Thinking Mode)	87.08%	30.83%	59.0%
GPT-4.1	17.50%	100.00%	58.8%
Granite-4.0-H-Micro	91.25%	19.58%	55.4%
Qwen2.5-7B-Instruct	51.25%	47.50%	49.4%
Granite-4.0-H-Tiny	82.92%	5.42%	44.2%
Gemini-2.0-Flashlite	61.67%	24.17%	42.9%
GPT-5-Nano (Medium Reasoning)	32.92%	48.75%	40.8%
Nova-Micro	15.24%	44.76%	30.0%
Nova-2-lite	21.67%	10.00%	15.8%
Llama-3.1-8B-Instruct	5.56%	5.56%	5.6%

Text Search and Extraction Results

Model	Q301	Q302	Q303	Q304	Average
GPT-5 (Medium Reasoning)	100.00%	99.58%	93.33%	87.50%	95.10%
GPT-5.2 (Medium Reasoning)	92.08%	87.08%	97.50%	95.42%	93.02%
Claude-Sonnet-4.5	100.00%	100.00%	85.83%	85.42%	92.81%
Kimi-K2-Thinking	98.33%	93.33%	84.44%	82.22%	89.58%
Qwen-Plus-2025-09-11	100.00%	73.75%	90.42%	92.08%	89.06%
Qwen3-Next-80B-A3B-Instruct	99.58%	68.33%	92.92%	91.67%	88.13%
DeepSeek-v3.1	96.25%	99.58%	85.83%	70.00%	87.92%
Qwen3-Max-2025-09-23	100.00%	100.00%	89.58%	61.25%	87.71%
Qwen3-235B-A22B-Instruct-2507	99.58%	100.00%	73.33%	77.92%	87.71%
Gemini-2.5-Pro	100.00%	100.00%	92.08%	52.92%	86.25%
Gemini-3-Pro-Preview	100.00%	99.58%	84.58%	60.83%	86.25%
Qwen3-Coder-480B-A35B-Instruct	100.00%	100.00%	86.25%	58.33%	86.15%
MiniMax-M2	98.75%	93.33%	85.83%	55.42%	83.33%
Nova-Premier	100.00%	100.00%	71.25%	57.92%	82.29%
GPT-5-Mini (Medium Reasoning)	97.92%	99.17%	78.75%	52.50%	82.09%
Claude-3.5-Haiku-20241022	97.78%	70.00%	68.89%	88.89%	81.39%
Qwen2.5-72B-Instruct	99.05%	91.43%	60.48%	69.52%	80.12%
Qwen3-30B-A3B (Thinking Mode)	97.08%	82.92%	46.25%	87.50%	78.44%
Mistral-Large-3-675B-Instruct-2512	97.50%	96.67%	54.58%	65.00%	78.44%
GLM-4.5	100.00%	100.00%	67.08%	40.83%	76.98%
Claude-Haiku-4.5	97.92%	90.83%	70.42%	45.00%	76.04%
GPT-5.1 (Medium Reasoning)	83.33%	67.50%	75.00%	78.33%	76.04%
GLM-4.6	100.00%	99.58%	67.08%	33.75%	75.10%
Gemini-2.0-Flash	94.58%	83.33%	58.33%	60.42%	74.17%
Qwen3-235B-A22B-Instruct-2507-FP8	99.58%	100.00%	40.00%	55.00%	73.65%

Llama-3.3-70B-Instruct	99.58%	99.17%	39.17%	56.25%	73.54%
Claude-Haiku-3.5	96.00%	72.00%	55.33%	70.67%	73.50%
MiniMax-M2.1	97.50%	94.58%	47.92%	52.92%	73.23%
Llama-3.1-70B-Instruct	98.75%	83.75%	55.00%	54.17%	72.92%
Gemini-2.5-Flash	100.00%	97.08%	60.00%	26.25%	70.83%
Llama-3.3-70B-Instruct-FP8-KV	99.17%	98.75%	50.42%	20.83%	67.29%
Llama-4-Scout-17B-16E-Instruct	97.50%	30.00%	80.00%	55.42%	65.73%
GLM-4.5-Air	100.00%	99.33%	55.33%	6.00%	65.17%
Qwen2.5-14B-Instruct	96.67%	78.33%	37.08%	21.25%	58.33%
Qwen3-32B (Thinking Mode)	97.50%	50.83%	33.75%	51.25%	58.33%
Gemini-2.5-Flash-Lite	99.17%	90.83%	3.33%	38.33%	57.92%
Qwen3-14B (Thinking Mode)	99.58%	66.25%	42.92%	17.50%	56.56%
Llama-4-Maverick-17B-128E-Instruct	97.08%	68.75%	39.17%	10.42%	53.86%
Llama-4-Maverick-17B-128E-Instruct-FP8	99.17%	62.08%	37.08%	8.75%	51.77%
Qwen2.5-32B-Instruct	98.75%	98.33%	1.67%	3.33%	50.52%
DeepSeek-v3	100.00%	97.50%	0.83%	2.08%	50.10%
Qwen3-14B-FP8	100.00%	95.83%	0.00%	0.00%	48.96%
Qwen3-14B	100.00%	93.75%	0.00%	0.00%	48.44%
Qwen3-32B-FP8	99.17%	92.92%	0.00%	0.00%	48.02%
Qwen3-32B	99.17%	89.17%	0.00%	0.00%	47.09%
Qwen3-Max-Preview	96.25%	87.08%	0.00%	0.00%	45.83%
Phi-4	86.19%	88.10%	3.33%	0.00%	44.41%
GPT-4.1	100.00%	53.75%	7.08%	1.67%	40.63%
Granite-4.0-H-Small	86.67%	74.17%	0.00%	0.00%	40.21%
Mistral-Large-Instruct-2411	99.58%	59.17%	0.83%	0.00%	39.90%
Qwen3-30B-A3B	95.00%	63.75%	0.00%	0.00%	39.69%
Qwen3-4B-Instruct-2507	64.17%	80.00%	1.25%	0.00%	36.36%
Qwen3-8B (Thinking Mode)	80.42%	34.58%	20.00%	8.75%	35.94%

Nova-2-lite	68.33%	70.00%	0.00%	0.00%	34.58%
Nova-Pro	100.00%	28.75%	1.67%	0.83%	32.81%
Qwen3-4B (Thinking Mode)	87.50%	31.25%	10.42%	0.83%	32.50%
Qwen3-235B-A22B	93.33%	11.25%	6.25%	13.33%	31.04%
Gemini-2.0-Flashlite	45.42%	55.00%	7.92%	11.67%	30.00%
Nova-Micro	50.48%	10.48%	30.95%	25.24%	29.29%
Qwen3-4B	57.78%	40.00%	0.00%	0.00%	24.45%
Granite-4.0-H-Tiny	81.67%	15.83%	0.00%	0.00%	24.38%
GPT-5.1	72.08%	20.00%	0.42%	1.67%	23.54%
Nova-Lite	85.42%	5.83%	0.00%	0.42%	22.92%
Qwen3-8B	62.00%	25.33%	0.00%	0.00%	21.83%
Qwen3-30B-A3B-Instruct-2507	27.08%	42.08%	14.17%	3.75%	21.77%
GPT-5-Nano (Medium Reasoning)	34.58%	13.33%	29.17%	7.08%	21.04%
Qwen-Flash-2025-07-28	23.75%	20.42%	12.92%	20.42%	19.38%
Llama-3.1-8B-Instruct	10.00%	12.22%	0.56%	0.00%	5.70%
Granite-4.0-H-Micro	8.33%	1.67%	0.00%	0.00%	2.50%
Qwen2.5-7B-Instruct	2.08%	1.25%	0.00%	0.00%	0.83%

CSV Processing Results

Model	Q401	Q402	Q403	Average
MiniMax-M2	100.00%	94.58%	98.33%	97.6%
GLM-4.6	92.92%	97.92%	100.00%	96.9%
Qwen3-Coder-480B-A35B-Instruct	100.00%	90.00%	99.58%	96.5%
Qwen3-Max-2025-09-23	99.58%	88.33%	99.58%	95.8%
Qwen3-235B-A22B-Instruct-2507-FP8	94.17%	94.17%	95.42%	94.6%
DeepSeek-v3.1	100.00%	80.83%	98.75%	93.2%
Claude-Sonnet-4.5	100.00%	77.50%	100.00%	92.5%
Gemini-2.5-Pro	98.75%	78.33%	98.33%	91.8%
Qwen3-235B-A22B-Instruct-2507	98.75%	77.08%	97.50%	91.1%
Gemini-3-Pro-Preview	100.00%	71.67%	100.00%	90.6%
MiniMax-M2.1	90.00%	82.50%	96.25%	89.6%
GPT-5 (Medium Reasoning)	97.92%	62.08%	99.58%	86.5%
GPT-5.2 (Medium Reasoning)	97.50%	60.00%	90.83%	82.8%
Kimi-K2-Thinking	98.89%	54.44%	85.00%	79.4%
Gemini-2.5-Flash	96.25%	38.33%	86.67%	73.8%
GLM-4.5	70.83%	50.83%	93.33%	71.7%
Mistral-Large-3-675B-Instruct-2512	78.75%	53.33%	79.58%	70.6%
Qwen3-Next-80B-A3B-Instruct	99.17%	14.17%	96.25%	69.9%
Qwen2.5-72B-Instruct	83.81%	40.95%	81.43%	68.7%
Qwen-Plus-2025-09-11	91.67%	10.83%	96.67%	66.4%
Qwen3-30B-A3B-Instruct-2507	75.83%	70.42%	50.42%	65.6%
Claude-3.5-Haiku-20241022	100.00%	4.44%	88.89%	64.4%
Qwen-Flash-2025-07-28	88.33%	42.08%	62.50%	64.3%
Claude-Haiku-3.5	98.67%	7.33%	86.67%	64.2%
GLM-4.5-Air	38.67%	68.00%	82.67%	63.1%

GPT-5.1 (Medium Reasoning)	70.00%	38.33%	80.00%	62.8%
GPT-5-Mini (Medium Reasoning)	88.75%	35.83%	60.42%	61.7%
Qwen3-235B-A22B	91.25%	52.92%	32.92%	59.0%
Llama-4-Maverick-17B-128E-Instruct	83.75%	10.42%	67.08%	53.8%
Llama-3.3-70B-Instruct-FP8-KV	80.42%	25.42%	51.67%	52.5%
Qwen3-32B-FP8	59.17%	36.67%	58.33%	51.4%
Llama-3.3-70B-Instruct	77.92%	16.25%	59.17%	51.1%
Llama-4-Maverick-17B-128E-Instruct-FP8	80.83%	8.75%	59.58%	49.7%
Nova-Lite	96.25%	3.75%	47.92%	49.3%
Qwen3-14B (Thinking Mode)	64.17%	28.75%	46.67%	46.5%
Qwen3-8B (Thinking Mode)	55.83%	32.08%	43.75%	43.9%
Qwen3-Max-Preview	32.08%	6.67%	86.25%	41.7%
Qwen3-30B-A3B (Thinking Mode)	82.08%	0.00%	38.75%	40.3%
Qwen3-32B	64.58%	46.25%	9.17%	40.0%
Qwen3-32B (Thinking Mode)	53.75%	13.75%	50.83%	39.4%
Qwen2.5-14B-Instruct	63.75%	20.42%	25.42%	36.5%
Mistral-Large-Instruct-2411	76.25%	3.75%	28.75%	36.3%
Llama-3.1-70B-Instruct	55.00%	39.58%	9.17%	34.6%
Llama-4-Scout-17B-16E-Instruct	67.92%	0.00%	30.42%	32.8%
Qwen2.5-32B-Instruct	26.25%	20.42%	40.83%	29.2%
Nova-Micro	47.62%	1.90%	29.52%	26.3%
Gemini-2.0-Flash	50.83%	5.83%	20.83%	25.8%
Nova-2-lite	71.25%	0.00%	1.25%	24.2%
Nova-Pro	8.75%	42.92%	20.00%	23.9%
DeepSeek-v3	18.75%	12.50%	21.67%	17.6%
Qwen2.5-7B-Instruct	29.17%	1.67%	20.83%	17.2%
Gemini-2.0-Flashlite	13.33%	4.58%	32.50%	16.8%
Claude-Haiku-4.5	33.75%	9.17%	6.67%	16.5%

Qwen3-4B (Thinking Mode)	13.33%	1.67%	30.83%	15.3%
Nova-Premier	32.50%	0.00%	10.00%	14.2%
GPT-5.1	7.08%	1.25%	27.08%	11.8%
GPT-5-Nano (Medium Reasoning)	15.00%	1.67%	11.67%	9.4%
Gemini-2.5-Flash-Lite	4.17%	2.50%	13.33%	6.7%
Qwen3-4B-Instruct-2507	2.50%	1.25%	15.83%	6.5%
Llama-3.1-8B-Instruct	7.22%	1.11%	10.00%	6.1%
GPT-4.1	9.17%	1.67%	5.83%	5.6%
Granite-4.0-H-Small	0.00%	0.83%	9.58%	3.5%
Qwen3-8B	5.33%	1.33%	1.33%	2.7%
Qwen3-4B	0.00%	0.00%	6.67%	2.2%
Granite-4.0-H-Tiny	4.58%	0.00%	0.42%	1.7%
Phi-4	0.00%	0.00%	3.81%	1.3%
Qwen3-30B-A3B	0.42%	0.00%	1.67%	0.7%
Qwen3-14B-FP8	0.00%	0.00%	0.83%	0.3%
Qwen3-14B	0.42%	0.00%	0.00%	0.1%
Granite-4.0-H-Micro	0.00%	0.00%	0.00%	0.0%

Database Processing Results

Model	Q501	Q502	Q503	Average
GPT-5 (Medium Reasoning)	95.00%	90.00%	100.00%	95.0%
GLM-4.6	100.00%	70.00%	100.00%	90.0%
Claude-Sonnet-4.5	90.00%	75.00%	100.00%	88.3%
MiniMax-M2	99.17%	65.00%	99.17%	87.8%
GLM-4.5	100.00%	63.33%	99.17%	87.5%
MiniMax-M2.1	87.50%	69.58%	100.00%	85.7%
Gemini-3-Pro-Preview	100.00%	54.58%	100.00%	84.9%
GPT-5-Mini (Medium Reasoning)	96.67%	92.50%	56.25%	81.8%
Qwen3-Coder-480B-A35B-Instruct	97.92%	47.08%	99.17%	81.4%
Qwen3-235B-A22B-Instruct-2507-FP8	70.42%	75.42%	90.83%	78.9%
Kimi-K2-Thinking	90.00%	51.67%	93.33%	78.3%
DeepSeek-v3.1	100.00%	52.92%	80.83%	77.9%
GLM-4.5-Air	82.67%	52.00%	96.67%	77.1%
Qwen3-235B-A22B-Instruct-2507	73.75%	65.00%	87.50%	75.4%
Qwen3-Max-Preview	62.08%	65.42%	98.33%	75.3%
Qwen3-Next-80B-A3B-Instruct	97.92%	22.92%	99.58%	73.5%
Qwen-Plus-2025-09-11	98.33%	21.25%	100.00%	73.2%
Gemini-2.5-Pro	86.67%	29.58%	96.67%	71.0%
GPT-5.2 (Medium Reasoning)	53.33%	57.50%	94.58%	68.5%
Llama-3.3-70B-Instruct	51.25%	65.00%	87.92%	68.1%
Qwen3-30B-A3B-Instruct-2507	84.58%	23.33%	93.75%	67.2%
Llama-4-Maverick-17B-128E-Instruct	97.08%	48.33%	54.58%	66.7%
Mistral-Large-3-675B-Instruct-2512	95.83%	3.75%	100.00%	66.5%
Nova-Premier	95.00%	5.00%	98.33%	66.1%
Gemini-2.5-Flash	95.00%	5.42%	97.08%	65.8%

Llama-4-Maverick-17B-128E-Instruct-FP8	95.42%	42.50%	52.08%	63.3%
Qwen3-4B-Instruct-2507	72.92%	22.92%	92.08%	62.6%
Phi-4	85.71%	18.57%	83.33%	62.5%
Nova-Pro	53.33%	54.17%	77.92%	61.8%
Qwen3-14B (Thinking Mode)	85.42%	5.83%	91.67%	61.0%
Qwen3-32B (Thinking Mode)	87.92%	4.58%	90.42%	61.0%
Nova-Lite	63.33%	55.83%	61.25%	60.1%
Qwen-Flash-2025-07-28	95.83%	22.50%	60.42%	59.6%
Llama-3.1-70B-Instruct	82.08%	12.50%	78.33%	57.6%
Qwen3-8B (Thinking Mode)	63.33%	23.33%	85.83%	57.5%
GPT-4.1	20.83%	68.33%	81.25%	56.8%
DeepSeek-v3	53.33%	27.50%	88.33%	56.4%
Mistral-Large-Instruct-2411	95.83%	0.83%	70.83%	55.8%
Llama-3.3-70B-Instruct-FP8-KV	67.50%	15.00%	82.50%	55.0%
Qwen3-Max-2025-09-23	18.33%	50.00%	95.00%	54.4%
Qwen3-32B-FP8	78.33%	2.50%	71.67%	50.8%
GPT-5.1 (Medium Reasoning)	35.83%	39.17%	75.00%	50.0%
Qwen3-32B	80.00%	3.75%	64.58%	49.4%
Qwen3-30B-A3B (Thinking Mode)	62.50%	20.83%	63.75%	49.0%
Nova-Micro	53.81%	36.19%	51.43%	47.1%
Qwen3-235B-A22B	62.08%	45.83%	32.92%	46.9%
Claude-3.5-Haiku-20241022	86.67%	3.33%	47.78%	45.9%
Qwen3-14B-FP8	33.33%	1.67%	100.00%	45.0%
Gemini-2.0-Flashlite	40.83%	2.50%	87.08%	43.5%
Claude-Haiku-3.5	82.00%	2.67%	44.67%	43.1%
Granite-4.0-H-Small	50.42%	20.83%	57.08%	42.8%
GPT-5.1	17.92%	22.50%	83.75%	41.4%
Qwen2.5-14B-Instruct	66.67%	15.83%	40.83%	41.1%

Qwen2.5-72B-Instruct	87.14%	2.38%	33.33%	41.0%
Qwen3-14B	19.58%	0.42%	100.00%	40.0%
Qwen3-30B-A3B	65.42%	33.33%	16.25%	38.3%
Gemini-2.5-Flash-Lite	50.00%	2.08%	60.00%	37.4%
Qwen3-4B (Thinking Mode)	52.50%	0.83%	55.83%	36.4%
Claude-Haiku-4.5	11.67%	18.75%	58.75%	29.7%
Llama-4-Scout-17B-16E-Instruct	21.25%	1.67%	64.58%	29.2%
Qwen2.5-7B-Instruct	18.33%	1.67%	39.58%	19.9%
Qwen2.5-32B-Instruct	2.92%	6.67%	46.25%	18.6%
Qwen3-8B	36.67%	0.67%	16.00%	17.8%
Gemini-2.0-Flash	37.92%	7.50%	3.33%	16.3%
GPT-5-Nano (Medium Reasoning)	5.83%	5.83%	22.08%	11.2%
Granite-4.0-H-Micro	0.00%	0.00%	13.75%	4.6%
Qwen3-4B	3.33%	0.00%	6.67%	3.3%
Granite-4.0-H-Tiny	0.00%	0.00%	7.92%	2.6%
Llama-3.1-8B-Instruct	2.22%	0.00%	5.00%	2.4%
Nova-2-lite	0.83%	0.00%	0.42%	0.4%

Database Processing (Guided) Results

Model	Q601	Q602	Average
GLM-4.6	100.00%	97.50%	98.8%
GPT-5 (Medium Reasoning)	100.00%	94.17%	97.1%
Claude-Sonnet-4.5	94.58%	99.58%	97.1%
Qwen3-Coder-480B-A35B-Instruct	100.00%	93.75%	96.9%
MiniMax-M2	97.08%	94.58%	95.8%
GPT-5-Mini (Medium Reasoning)	98.33%	93.33%	95.8%
GLM-4.5	98.33%	92.08%	95.2%
DeepSeek-v3.1	99.58%	87.50%	93.5%
GPT-5.2 (Medium Reasoning)	97.08%	86.25%	91.7%
Gemini-3-Pro-Preview	100.00%	80.83%	90.4%
Qwen3-Max-Preview	100.00%	75.42%	87.7%
Qwen3-235B-A22B-Instruct-2507-FP8	100.00%	71.67%	85.8%
Qwen3-Max-2025-09-23	99.17%	70.00%	84.6%
Qwen3-235B-A22B-Instruct-2507	100.00%	67.92%	84.0%
Gemini-2.5-Pro	99.58%	62.92%	81.3%
GLM-4.5-Air	83.33%	76.67%	80.0%
Nova-Pro	100.00%	59.17%	79.6%
Qwen3-235B-A22B	100.00%	55.00%	77.5%
GPT-5.1 (Medium Reasoning)	95.83%	58.33%	77.1%
MiniMax-M2.1	100.00%	53.75%	76.9%
Qwen-Flash-2025-07-28	100.00%	52.50%	76.3%
Llama-4-Maverick-17B-128E-Instruct-FP8	100.00%	45.83%	72.9%
Nova-Premier	100.00%	44.17%	72.1%
Qwen3-30B-A3B-Instruct-2507	100.00%	43.75%	71.9%
Llama-4-Maverick-17B-128E-Instruct	100.00%	43.33%	71.7%

Kimi-K2-Thinking	89.44%	50.56%	70.0%
GPT-5.1	86.67%	50.83%	68.8%
Llama-3.3-70B-Instruct	100.00%	35.83%	67.9%
Llama-4-Scout-17B-16E-Instruct	99.58%	35.42%	67.5%
Qwen3-30B-A3B	100.00%	30.42%	65.2%
Llama-3.1-70B-Instruct	100.00%	30.00%	65.0%
Qwen3-8B (Thinking Mode)	98.33%	31.67%	65.0%
Granite-4.0-H-Small	100.00%	25.83%	62.9%
Llama-3.3-70B-Instruct-FP8-KV	100.00%	24.58%	62.3%
Phi-4	99.52%	24.76%	62.1%
Qwen3-30B-A3B (Thinking Mode)	100.00%	20.83%	60.4%
Qwen-Plus-2025-09-11	98.75%	20.83%	59.8%
Qwen2.5-32B-Instruct	100.00%	16.67%	58.3%
Qwen3-4B-Instruct-2507	100.00%	14.58%	57.3%
Qwen3-Next-80B-A3B-Instruct	98.75%	15.42%	57.1%
Mistral-Large-3-675B-Instruct-2512	100.00%	13.75%	56.9%
Qwen3-32B	100.00%	13.75%	56.9%
Qwen3-32B-FP8	100.00%	12.92%	56.5%
Qwen2.5-14B-Instruct	100.00%	12.50%	56.3%
Qwen3-14B-FP8	100.00%	12.08%	56.0%
Qwen3-32B (Thinking Mode)	100.00%	11.67%	55.8%
Nova-Lite	100.00%	8.33%	54.2%
Nova-Micro	100.00%	8.10%	54.1%
DeepSeek-v3	100.00%	6.67%	53.3%
Qwen3-14B	100.00%	6.25%	53.1%
Qwen3-8B	100.00%	6.00%	53.0%
Gemini-2.0-Flash	90.00%	15.42%	52.7%
Gemini-2.5-Flash	100.00%	5.00%	52.5%

Qwen3-14B (Thinking Mode)	100.00%	5.00%	52.5%
Mistral-Large-Instruct-2411	99.58%	2.92%	51.3%
Gemini-2.5-Flash-Lite	100.00%	1.67%	50.8%
Qwen2.5-72B-Instruct	100.00%	0.95%	50.5%
Claude-3.5-Haiku-20241022	94.44%	4.44%	49.4%
Gemini-2.0-Flashlite	96.25%	1.67%	49.0%
Claude-Haiku-3.5	96.00%	1.33%	48.7%
Qwen2.5-7B-Instruct	97.08%	0.00%	48.5%
Qwen3-4B (Thinking Mode)	94.58%	0.42%	47.5%
Qwen3-4B	86.67%	0.00%	43.3%
GPT-4.1	66.67%	7.92%	37.3%
Granite-4.0-H-Tiny	61.25%	0.00%	30.6%
GPT-5-Nano (Medium Reasoning)	8.75%	4.17%	6.5%
Llama-3.1-8B-Instruct	11.11%	0.00%	5.6%
Claude-Haiku-4.5	2.92%	7.50%	5.2%
Nova-2-lite	2.92%	0.00%	1.5%
Granite-4.0-H-Micro	0.00%	0.00%	0.0%

Instruction Following Results

Model	Q701	Q702	Q703	Avg
GPT-5 (Medium Reasoning)	100.00%	100.00%	100.00%	100.0%
GLM-4.6	100.00%	100.00%	100.00%	100.0%
DeepSeek-v3.1	100.00%	100.00%	100.00%	100.0%
Qwen3-235B-A22B-Instruct-2507	100.00%	100.00%	100.00%	100.0%
Qwen3-235B-A22B-Instruct-2507-FP8	100.00%	100.00%	100.00%	100.0%
Llama-3.3-70B-Instruct	100.00%	100.00%	100.00%	100.0%
Llama-3.3-70B-Instruct-FP8-KV	100.00%	100.00%	100.00%	100.0%
Qwen3-Max-Preview	100.00%	100.00%	100.00%	100.0%
Llama-4-Maverick-17B-128E-Instruct-FP8	100.00%	100.00%	100.00%	100.0%
Qwen3-30B-A3B (Thinking Mode)	100.00%	100.00%	100.00%	100.0%
Qwen3-30B-A3B-Instruct-2507	100.00%	100.00%	100.00%	100.0%
Qwen3-14B (Thinking Mode)	100.00%	100.00%	100.00%	100.0%
Qwen-Flash-2025-07-28	100.00%	100.00%	100.00%	100.0%
Qwen3-235B-A22B	100.00%	100.00%	100.00%	100.0%
Qwen2.5-14B-Instruct	100.00%	100.00%	100.00%	100.0%
Llama-4-Scout-17B-16E-Instruct	100.00%	100.00%	100.00%	100.0%
Qwen3-4B-Instruct-2507	100.00%	100.00%	100.00%	100.0%
Qwen3-14B	100.00%	100.00%	100.00%	100.0%
Qwen3-30B-A3B	100.00%	100.00%	100.00%	100.0%
GLM-4.5	99.58%	100.00%	100.00%	99.9%
Qwen3-14B-FP8	100.00%	99.58%	100.00%	99.9%
Llama-4-Maverick-17B-128E-Instruct	100.00%	99.17%	100.00%	99.7%
Qwen-Plus-2025-09-11	99.58%	98.75%	100.00%	99.4%
Llama-3.1-70B-Instruct	100.00%	100.00%	97.50%	99.2%
Qwen3-8B	100.00%	97.33%	100.00%	99.1%

Qwen3-Next-80B-A3B-Instruct	99.17%	97.92%	100.00%	99.0%
Qwen3-Max-2025-09-23	99.58%	98.33%	97.92%	98.6%
Granite-4.0-H-Small	100.00%	95.42%	100.00%	98.5%
Claude-Haiku-3.5	94.67%	99.33%	99.33%	97.8%
GPT-5-Mini (Medium Reasoning)	98.75%	93.33%	97.50%	96.5%
Claude-3.5-Haiku-20241022	91.11%	96.67%	97.78%	95.2%
GPT-5.2 (Medium Reasoning)	96.25%	93.75%	95.42%	95.1%
Qwen2.5-7B-Instruct	97.50%	86.25%	95.42%	93.1%
Qwen3-Coder-480B-A35B-Instruct	100.00%	75.83%	100.00%	91.9%
GPT-5.1 (Medium Reasoning)	93.33%	90.00%	80.83%	88.1%
Qwen3-4B (Thinking Mode)	92.92%	86.67%	84.17%	87.9%
MiniMax-M2.1	100.00%	63.75%	100.00%	87.9%
Gemini-2.5-Flash	100.00%	61.25%	99.58%	86.9%
Qwen3-32B (Thinking Mode)	100.00%	58.75%	100.00%	86.3%
Nova-Lite	100.00%	57.50%	100.00%	85.8%
Qwen3-8B (Thinking Mode)	97.92%	46.25%	97.92%	80.7%
MiniMax-M2	98.33%	30.00%	100.00%	76.1%
GPT-4.1	95.83%	27.08%	100.00%	74.3%
Gemini-3-Pro-Preview	100.00%	17.92%	100.00%	72.6%
Nova-Pro	100.00%	3.33%	100.00%	67.8%
Mistral-Large-Instruct-2411	99.58%	2.50%	100.00%	67.4%
Qwen3-32B-FP8	100.00%	0.42%	100.00%	66.8%
Gemini-2.5-Pro	100.00%	0.00%	100.00%	66.7%
Mistral-Large-3-675B-Instruct-2512	100.00%	0.00%	100.00%	66.7%
Qwen2.5-72B-Instruct	100.00%	0.00%	100.00%	66.7%
Nova-Premier	100.00%	0.00%	100.00%	66.7%
Qwen3-32B	100.00%	0.00%	100.00%	66.7%
DeepSeek-v3	100.00%	0.00%	100.00%	66.7%

Qwen2.5-32B-Instruct	100.00%	0.00%	100.00%	66.7%
Gemini-2.5-Flash-Lite	93.33%	3.75%	100.00%	65.7%
Claude-Sonnet-4.5	98.33%	0.00%	97.08%	65.1%
Phi-4	99.52%	0.00%	95.24%	64.9%
GLM-4.5-Air	92.67%	0.67%	100.00%	64.4%
Kimi-K2-Thinking	93.89%	38.33%	58.89%	63.7%
Gemini-2.0-Flashlite	97.92%	0.00%	85.83%	61.3%
Gemini-2.0-Flash	87.08%	0.00%	93.75%	60.3%
Nova-Micro	95.71%	0.00%	77.14%	57.6%
GPT-5.1	90.83%	48.33%	14.58%	51.2%
Qwen3-4B	32.22%	42.22%	72.22%	48.9%
GPT-5-Nano (Medium Reasoning)	21.25%	54.58%	32.50%	36.1%
Granite-4.0-H-Tiny	46.25%	16.25%	31.67%	31.4%
Nova-2-lite	8.33%	0.00%	79.17%	29.2%
Llama-3.1-8B-Instruct	12.78%	11.11%	50.00%	24.6%
Claude-Haiku-4.5	6.25%	0.00%	3.33%	3.2%
Granite-4.0-H-Micro	0.00%	0.00%	0.00%	0.0%

Important Information About this Report

CONTRIBUTORS
Mitch Lewis
Performance Analyst | Signal65

PUBLISHER
Ryan Shrout
President and GM | Signal65

INQUIRIES
Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS
This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

DISCLOSURES
Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT SIGNAL65
Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

CONTACT INFORMATION
Signal65 I signal65.com