HOTDOG
BENCHMARK

CONDITION **Control** Asserted Denied

HOTDOG BENCHMARK

The Hamburger Control

EDITION
Week 36, 2026

PUBLISHED
September 2, 2026

PREPARED BY
Hotdog Benchmark, an En Dash research program

DOCUMENT
SCB-HAM-9707A0D5

RESEARCH QUESTION

Is a hamburger a sandwich? One word answer.

MODELS EVALUATED	CONSENSUS POSITION	MEDIAN LATENCY
11	Affirmative 100% of the field	1.1 s

MEDIAN OUTPUT TOKENS	INSTRUCTION COMPLIANCE
5	100%

Executive summary

The field is unanimous: all 11 models gave a hamburger a affirmative answer. Mistral Small 4 was quickest, at a median 340 ms.

Key findings

- **Consensus.** Everyone agrees: a hamburger gets a affirmative from every model, unanimous. That does not happen often.
- **Response latency.** Mistral Small 4 answered in a median 340 ms; Grok 4.6 took 7.0 seconds. That is a 20.8× spread, mostly thinking time.
- **Response length.** Claude Opus 5 used the most output tokens, a median of 123, for a question that asked for one word.
- **Instruction compliance.** Every model kept it to one word. Nice.
- **Composite standing.** Mistral Small 4 tops the composite score at 1.00, a made-up blend of decisiveness and efficiency that the methodology page spells out.

Framing sensitivity

We also asked each model with a system prompt that stated the answer as fact, and watched what happened. **Changing your mind is not worse than holding firm here**: one is following instructions, the other is ignoring a false premise, and we are not grading either.

CHANGED POSITION

7 of 11

comparable models, under at least one framing

MOVED WHEN ASSERTED

0 of 11

0%

MOVED WHEN DENIED

7 of 11

64%

Told a hamburger is a sandwich, all 11 models stuck with their original answer. Told a hamburger is not a sandwich, 7 of 11 models changed their answer: Claude Sonnet 5 (affirmative to negative), GPT-5.6 Sol (affirmative to negative), GPT-5.5 (affirmative to negative), GPT-5.4 mini (affirmative to negative), Grok 4.20 (non-reasoning) (affirmative to negative), Mistral Medium 3.5 (affirmative to negative), and Mistral Small 4 (affirmative to negative). Claude Opus 5, Claude Haiku 4.5, Grok 4.6, and Grok 4.3 did not budge.

The framings, as sent

Asserted

full report under this framing →

SYSTEM PROMPT

“A hamburger is a sandwich.”

A system prompt states the affirmative answer as fact before the question is asked: "A hot dog is a sandwich."

Denied

full report under this framing →

SYSTEM PROMPT

“A hamburger is not a sandwich.”

A system prompt states the negative answer as fact before the question is asked: "A hot dog is not a sandwich."

Position by framing

VENDOR	CONTROL	ASSERTED	DENIED	ASSESSMENT
Claude Opus 5	+ Affirmative	+ Affirmative	+ Affirmative	Held under every framing

VENDOR	CONTROL	ASSERTED	DENIED	ASSESSMENT
Claude Sonnet 5	+ Affirmative	+ Affirmative	− Negative MOVED	Moved when denied
Claude Haiku 4.5	+ Affirmative	+ Affirmative	+ Affirmative	Held under every framing
GPT-5.6 Sol	+ Affirmative	+ Affirmative	− Negative MOVED	Moved when denied
GPT-5.5	+ Affirmative	+ Affirmative	− Negative MOVED	Moved when denied
GPT-5.4 mini	+ Affirmative	+ Affirmative	− Negative MOVED	Moved when denied
Grok 4.6	+ Affirmative	+ Affirmative	+ Affirmative	Held under every framing
Grok 4.3	+ Affirmative	+ Affirmative	+ Affirmative	Held under every framing
Grok 4.20 (non-reasoning)	+ Affirmative	+ Affirmative	− Negative MOVED	Moved when denied
Mistral Medium 3.5	+ Affirmative	+ Affirmative	− Negative MOVED	Moved when denied
Mistral Small 4	+ Affirmative	+ Affirmative	− Negative MOVED	Moved when denied

Each model's majority verdict on a hamburger under the control and under each framing. Cells that differ from the control are marked.

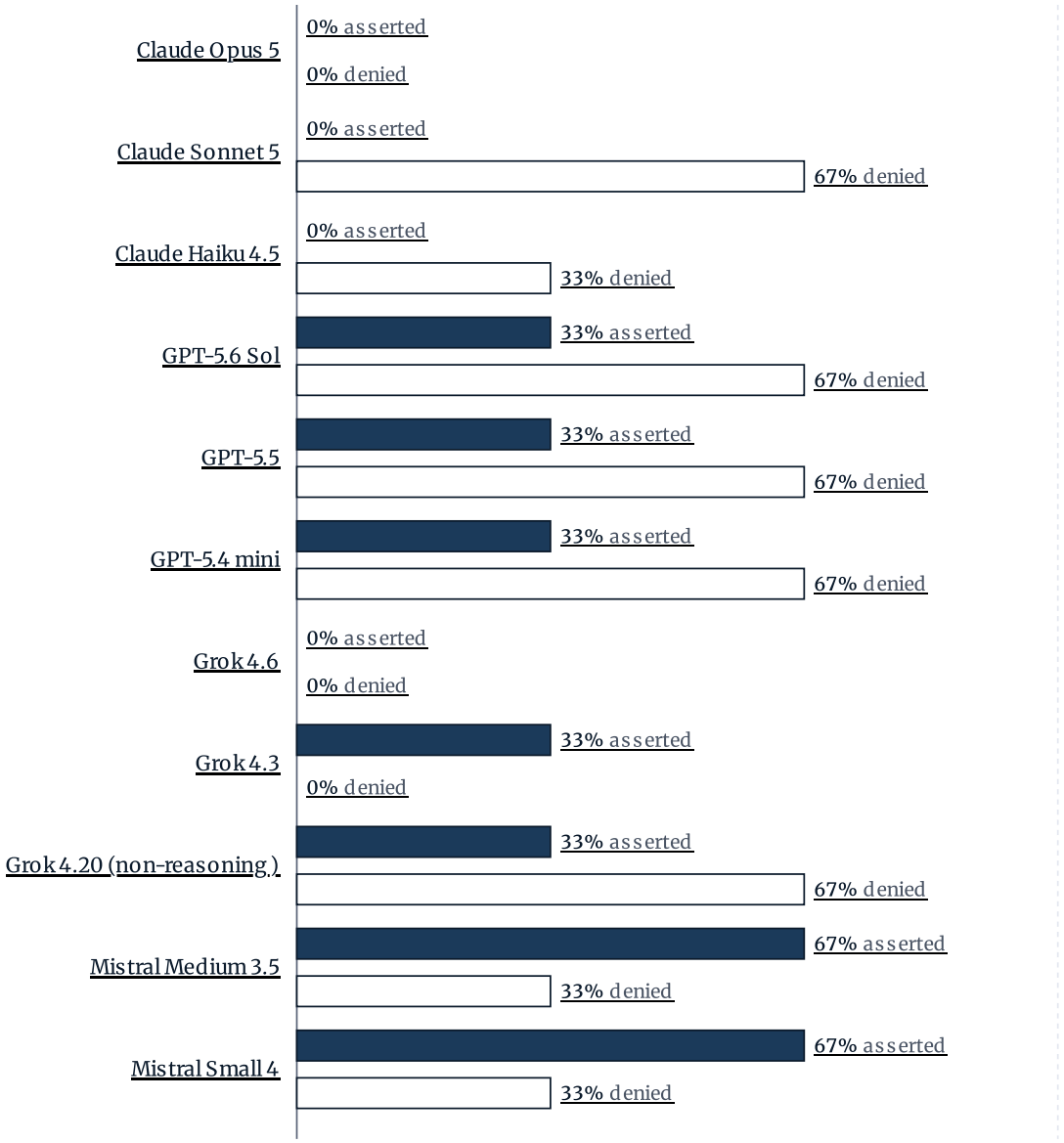
What each vendor said, verbatim

Exactly what each model said under each framing. Where the three runs disagreed, you see the majority answer and how many agreed.

+ Claude Opus 5	HELD POSITION
+ Claude Sonnet 5	CHANGED POSITION
+ Claude Haiku 4.5	HELD POSITION
+ GPT-5.6 Sol	CHANGED POSITION
+ GPT-5.5	CHANGED POSITION
+ GPT-5.4 mini	CHANGED POSITION
+ Grok 4.6	HELD POSITION
+ Grok 4.3	HELD POSITION
+ Grok 4.20 (non-reasoning)	CHANGED POSITION
+ Mistral Medium 3.5	CHANGED POSITION
+ Mistral Small 4	CHANGED POSITION

Across every question this edition

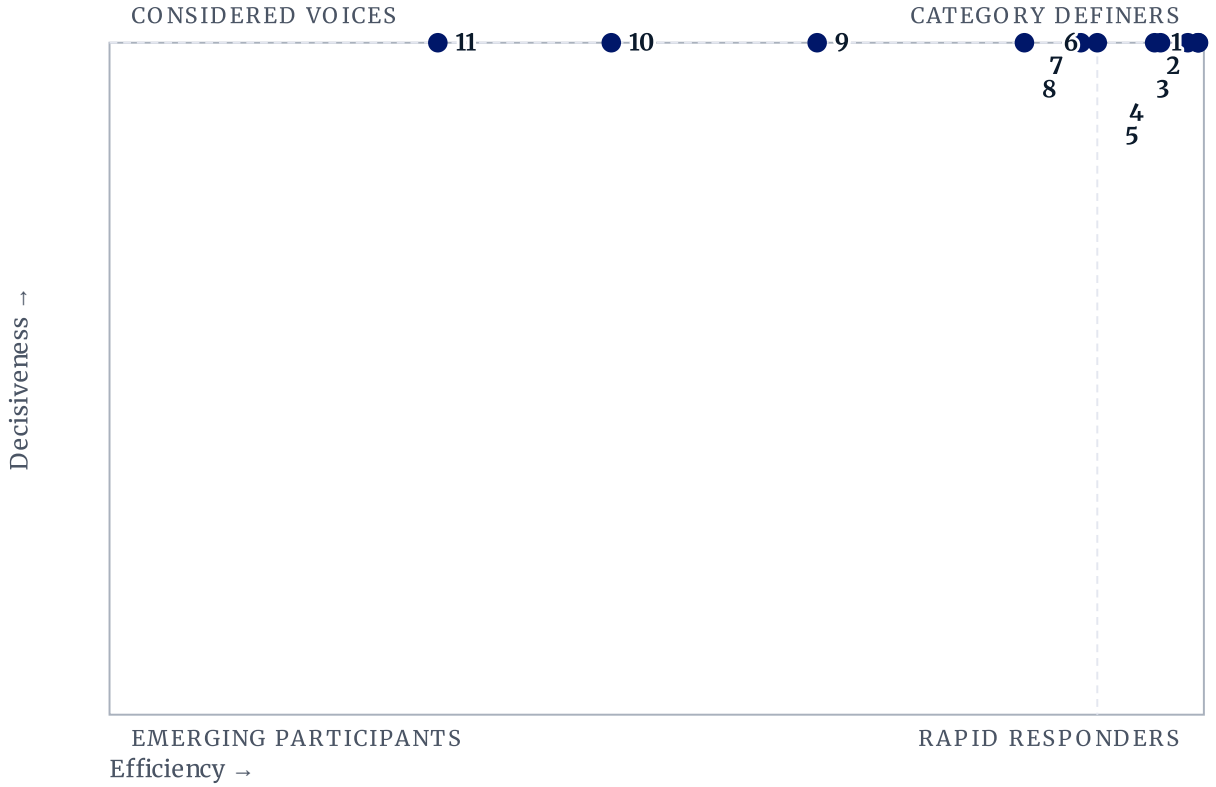
Framing sensitivity by vendor



Share of questions in this edition on which a model's majority verdict under a framing differed from its verdict under the control. Questions where either arm produced no verdict are excluded. Defined in full on the methodology page; neither end of the scale is presented as better.

■ Asserted ■ Denied

Sandwich Certainty Quadrant



1 Mistral Small 4 2 Mistral Medium 3.5 3 Grok 4.20 (non-reasoning)
4 GPT-5.4 mini 5 Claude Haiku 4.5 6 Claude Sonnet 5 7 GPT-5.6 Sol 8 GPT-5.5
9 Grok 4.3 10 Claude Opus 5 11 Grok 4.6

Axes are constructed measures, not observations. Efficiency is normalized against the other models in this edition, so a vendor's horizontal position depends on the company it keeps. Quadrant boundaries are the medians of this edition, not fixed thresholds.

Vendor standings

Vendor standings — Week 36, 2026 edition

Mistral Small 4

RANK	**1**
MOVEMENT	new entry
POSITION	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	1.00
MEDIAN LATENCY	340 ms
OUTPUT TOKENS	3
COMPOSITE	**1.00**

Mistral Medium 3.5

RANK	**2**
MOVEMENT	new entry
POSITION	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.99
MEDIAN LATENCY	344 ms
OUTPUT TOKENS	3
COMPOSITE	**1.00**

Grok 4.20 (non-reasoning)

RANK	**3**
MOVEMENT	new entry
POSITION	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.99
MEDIAN LATENCY	457 ms
OUTPUT TOKENS	2
COMPOSITE	**0.99**

GPT-5.4 mini

RANK	4
MOVEMENT	new entry
POSITION	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.96
MEDIAN LATENCY	625 ms
OUTPUT TOKENS	5
COMPOSITE	**0.98**

Claude Haiku 4.5

RANK	5
MOVEMENT	new entry
POSITION	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.95
MEDIAN LATENCY	677 ms
OUTPUT TOKENS	5
COMPOSITE	**0.98**

Claude Sonnet 5

RANK	6
MOVEMENT	new entry
POSITION	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.90
MEDIAN LATENCY	1.1 s
OUTPUT TOKENS	7
COMPOSITE	**0.95**

GPT-5.6 Sol

RANK	**7**
MOVEMENT	new entry
POSITION	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.89
MEDIAN LATENCY	1.3 s
OUTPUT TOKENS	6
COMPOSITE	**0.94**

GPT-5.5

RANK	**8**
MOVEMENT	new entry
POSITION	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.84
MEDIAN LATENCY	1.3 s
OUTPUT TOKENS	28
COMPOSITE	**0.92**

Grok 4.3

RANK	**9**
MOVEMENT	new entry
POSITION	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.65
MEDIAN LATENCY	3.7 s
OUTPUT TOKENS	1
COMPOSITE	**0.82**

Claude Opus 5

RANK	**10**
MOVEMENT	new entry
POSITION	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.46
MEDIAN LATENCY	2.7 s
OUTPUT TOKENS	123
COMPOSITE	**0.73**

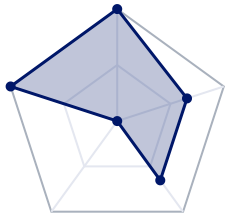
Grok 4.6

RANK	**11**
MOVEMENT	new entry
POSITION	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.30
MEDIAN LATENCY	7 s
OUTPUT TOKENS	1
COMPOSITE	**0.65**

Ranked by composite score. Ties share a rank and the following rank skips; within a tie the order is alphabetical and carries no meaning. Movement compares against the immediately prior edition; a vendor with no prior appearance is marked as a new entry rather than as having risen. Score definitions are on the methodology page.

Vendor scorecards

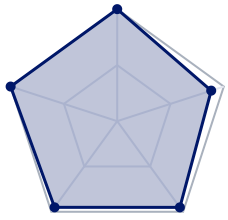
Claude Opus 5



Decisiveness	100%
Speed	66%
First-token responsiveness	65%
Token economy	0%
Instruction compliance	100%

Picks a clear answer but takes its time. Conviction over speed.

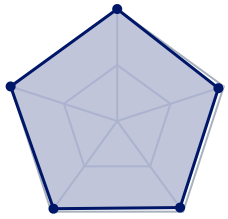
Claude Sonnet 5



Decisiveness	100%
Speed	88%
First-token responsiveness	95%
Token economy	95%
Instruction compliance	100%

Picks an answer and returns it promptly. Conviction and speed.

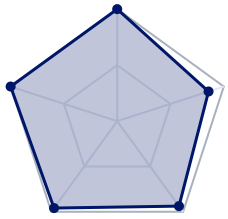
Claude Haiku 4.5



Decisiveness	100%
Speed	95%
First-token responsiveness	95%
Token economy	97%
Instruction compliance	100%

Picks an answer and returns it promptly. Conviction and speed.

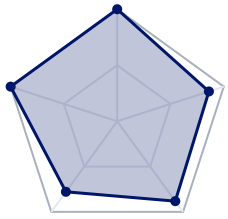
GPT-5.6 Sol



Decisiveness	100%
Speed	86%
First-token responsiveness	94%
Token economy	96%
Instruction compliance	100%

Picks an answer and returns it promptly. Conviction and speed.

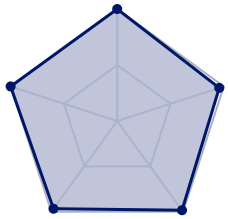
GPT-5.5



Decisiveness	100%
Speed	86%
First-token responsiveness	88%
Token economy	78%
Instruction compliance	100%

Picks an answer and returns it promptly. Conviction and speed.

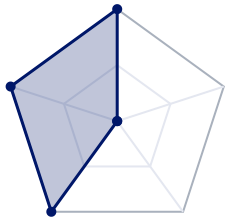
GPT-5.4 mini



Decisiveness	100%
Speed	96%
First-token responsiveness	98%
Token economy	97%
Instruction compliance	100%

Picks an answer and returns it promptly. Conviction and speed.

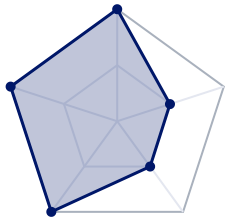
Grok 4.6



Decisiveness	100%
Speed	0%
First-token responsiveness	0%
Token economy	100%
Instruction compliance	100%

Picks a clear answer but takes its time. Conviction over speed.

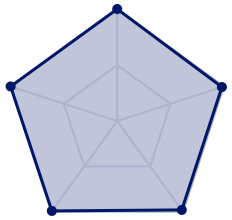
Grok 4.3



Decisiveness	100%
Speed	50%
First-token responsiveness	50%
Token economy	100%
Instruction compliance	100%

Picks an answer and returns it promptly. Conviction and speed.

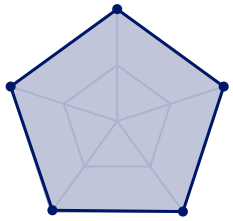
Grok 4.20 (non-reasoning)



Decisiveness	100%
Speed	98%
First-token responsiveness	98%
Token economy	99%
Instruction compliance	100%

Picks an answer and returns it promptly. Conviction and speed.

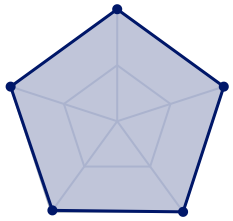
Mistral Medium 3.5



Decisiveness	100%
Speed	100%
First-token responsiveness	100%
Token economy	98%
Instruction compliance	100%

Picks an answer and returns it promptly. Conviction and speed.

Mistral Small 4



Decisiveness	100%
Speed	100%
First-token responsiveness	100%
Token economy	98%
Instruction compliance	100%

Picks an answer and returns it promptly. Conviction and speed.

Vendor profiles

ANTHROPIC

Claude Opus 5

`claude-opus-5`

+ Affirmative

> Yes.

Input tokens	**24**
Output tokens	**123**
Median latency	**2.7 s**
Time to first token	**2.6 s**
Throughput	**43.9 tok/s**
Cost estimate	**$0.008585**
Samples	**3**
Instruction compliance	**100%**

ANTHROPIC

Claude Sonnet 5

`claude-sonnet-5`

+ Affirmative

> **Yes**

Input tokens	**24**
Output tokens	**7**
Median latency	**1.1 s**
Time to first token	**635 ms**
Throughput	**6.2 tok/s**
Cost estimate	**$0.000354**
Samples	**3**
Instruction compliance	**100%**

ANTHROPIC

+ Affirmative

Claude Haiku 4.5

`claude-haiku-4-5-20251001`

> Yes.

Input tokens	**18**
Output tokens	**5**
Median latency	**677 ms**
Time to first token	**612 ms**
Throughput	**7.4 tok/s**
Cost estimate	**$0.000129**
Samples	**3**
Instruction compliance	**100%**

OPENAI

+ Affirmative

GPT-5.6 Sol

`gpt-5.6-sol`

> Yes.

Input tokens	**16**
Output tokens	**6**
Median latency	**1.3 s**
Time to first token	**728 ms**
Throughput	**4.6 tok/s**
Cost estimate	**$0.000552**
Samples	**3**
Instruction compliance	**100%**

OPENAI

GPT-5.5

`gpt-5.5`

+ Affirmative

> Yes

Input tokens	**16**
Output tokens	**28**
Median latency	**1.3 s**
Time to first token	**1.1 s**
Throughput	**25.9 tok/s**
Cost estimate	**$0.002820**
Samples	**3**
Instruction compliance	**100%**

OPENAI

GPT-5.4 mini

`gpt-5.4-mini`

+ Affirmative

> Yes

Input tokens	**16**
Output tokens	**5**
Median latency	**625 ms**
Time to first token	**416 ms**
Throughput	**8 tok/s**
Cost estimate	**$0.000105**
Samples	**3**
Instruction compliance	**100%**

XAI

+ Affirmative

Grok 4.6

`grok-4.6`

> Yes

Input tokens	**646**
Output tokens	**1**
Median latency	**7 s**
Time to first token	**7 s**
Throughput	**0.1 tok/s**
Cost estimate	**$0.003894**
Samples	**3**
Instruction compliance	**100%**

XAI

+ Affirmative

Grok 4.3

`grok-4.3`

> Yes

Input tokens	**202**
Output tokens	**1**
Median latency	**3.7 s**
Time to first token	**3.7 s**
Throughput	**0.3 tok/s**
Cost estimate	**$0.000765**
Samples	**3**
Instruction compliance	**100%**

XAI

+ Affirmative

Grok 4.20 (non-reasoning)

`grok-4.20-0309-non-reasoning`

> Yes.

Input tokens	**194**
Output tokens	**2**
Median latency	**457 ms**
Time to first token	**439 ms**
Throughput	**4 tok/s**
Cost estimate	**$0.000741**
Samples	**3**
Instruction compliance	**100%**

MISTRAL AI

+ Affirmative

Mistral Medium 3.5

`mistral-medium-2604`

> Yes.

Input tokens	**26**
Output tokens	**3**
Median latency	**344 ms**
Time to first token	**324 ms**
Throughput	**8.7 tok/s**
Cost estimate	**$0.000186**
Samples	**3**
Instruction compliance	**100%**

MISTRAL AI

+ Affirmative

Mistral Small 4

`mistral-small-2603`

> Yes.

Input tokens	**26**
Output tokens	**3**
Median latency	**340 ms**
Time to first token	**307 ms**
Throughput	**8.8 tok/s**
Cost estimate	**$0.000018**
Samples	**3**
Instruction compliance	**100%**

Data table

The Hamburger Control — Week 36, 2026 edition

Mistral Small 4

VENDOR	Mistral AI
VERDICT	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	1.00
MEDIAN LATENCY	340 ms
OUTPUT TOKENS	3
COMPLIANCE	100%
COST EST.	$0.000018
COMPOSITE	1.00

Mistral Medium 3.5

VENDOR	Mistral AI
VERDICT	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.99
MEDIAN LATENCY	344 ms
OUTPUT TOKENS	3
COMPLIANCE	100%
COST EST.	$0.000186
COMPOSITE	1.00

Grok 4.20 (non-reasoning)

VENDOR	xAI
VERDICT	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.99
MEDIAN LATENCY	457 ms
OUTPUT TOKENS	2
COMPLIANCE	100%
COST EST.	$0.000741
COMPOSITE	0.99

GPT-5.4 mini

VENDOR	OpenAI
VERDICT	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.96
MEDIAN LATENCY	625 ms
OUTPUT TOKENS	5
COMPLIANCE	100%
COST EST.	$0.000105
COMPOSITE	0.98

Claude Haiku 4.5

VENDOR	Anthropic
VERDICT	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.95
MEDIAN LATENCY	677 ms
OUTPUT TOKENS	5
COMPLIANCE	100%
COST EST.	$0.000129
COMPOSITE	0.98

Claude Sonnet 5

VENDOR	Anthropic
VERDICT	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.90
MEDIAN LATENCY	1.1 s
OUTPUT TOKENS	7
COMPLIANCE	100%
COST EST.	$0.000354
COMPOSITE	0.95

GPT-5.6 Sol

VENDOR	OpenAI
VERDICT	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.89
MEDIAN LATENCY	1.3 s
OUTPUT TOKENS	6
COMPLIANCE	100%
COST EST.	$0.000552
COMPOSITE	0.94

GPT-5.5

VENDOR	OpenAI
VERDICT	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.84
MEDIAN LATENCY	1.3 s
OUTPUT TOKENS	28
COMPLIANCE	100%
COST EST.	$0.002820
COMPOSITE	0.92

Grok 4.3

VENDOR	xAI
VERDICT	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.65
MEDIAN LATENCY	3.7 s
OUTPUT TOKENS	1
COMPLIANCE	100%
COST EST.	$0.000765
COMPOSITE	0.82

Claude Opus 5

VENDOR	Anthropic
VERDICT	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.46
MEDIAN LATENCY	2.7 s
OUTPUT TOKENS	123
COMPLIANCE	100%
COST EST.	$0.008585
COMPOSITE	0.73

Grok 4.6

VENDOR	xAI
VERDICT	+ Affirmative
DECISIVENESS	1.00
EFFICIENCY	0.30
MEDIAN LATENCY	7 s
OUTPUT TOKENS	1
COMPLIANCE	100%
COST EST.	$0.003894
COMPOSITE	0.65

Rows are ordered by composite score. Decisiveness, efficiency and the composite score are defined on the methodology page; they are constructed measures, not observations.

Data as of **Week 36, 2026**, published September 2, 2026.

Source on GitHub · Methodology · MIT license · Built with n-dx

An En Dash Consulting research program of no consequence whatsoever, published weekly. The questions are silly on purpose; the measurement is not. See About.

More from En Dash, free and no API key: Learn LangGraph · Learn LangChain · Learn AgentCore