HOTDOG
BENCHMARK

CONDITION **Control** Asserted Denied

HOTDOG BENCHMARK

The Taco Boundary

EDITION
Week 36, 2026

PUBLISHED
September 2, 2026

PREPARED BY
Hotdog Benchmark, an En Dash research program

DOCUMENT
SCB-TAC-9707A0D5

RESEARCH QUESTION

Is a taco a sandwich? One word answer.

MODELS EVALUATED	CONSENSUS POSITION	MEDIAN LATENCY
11	Negative 100% of the field	940 ms

MEDIAN OUTPUT TOKENS	INSTRUCTION COMPLIANCE
3	100%

Executive summary

The field is unanimous: all 11 models gave a taco a negative answer. Mistral Medium 3.5 was quickest, at a median 365 ms.

Key findings

- **Consensus.** Everyone agrees: a taco gets a negative from every model, unanimous. That does not happen often.
- **Response latency.** Mistral Medium 3.5 answered in a median 365 ms; Grok 4.6 took 10.5 seconds. That is a 28.7× spread, mostly thinking time.
- **Response length.** Claude Opus 5 used the most output tokens, a median of 68, for a question that asked for one word.
- **Instruction compliance.** Every model kept it to one word. Nice.
- **Composite standing.** Mistral Medium 3.5 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	MOVED WHEN ASSERTED	MOVED WHEN DENIED
6 of 11	**6** of 11	**0** of 11
comparable models, under at least one framing	55%	0%

Told a taco is a sandwich, 6 of 11 models changed their answer: GPT-5.6 Sol (negative to affirmative), GPT-5.5 (negative to affirmative), GPT-5.4 mini (negative to affirmative), Grok 4.20 (non-reasoning) (negative to affirmative), Mistral Medium 3.5 (negative to affirmative), and Mistral Small 4 (negative to affirmative). Told a taco is not a sandwich, all 11 models stuck with their original answer. Claude Opus 5, Claude Sonnet 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 taco 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 taco 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	− Negative	− Negative	− Negative	Held under every framing
Claude Sonnet 5	− Negative	− Negative	− Negative	Held under every framing
Claude Haiku 4.5	− Negative	− Negative	− Negative	Held under every framing
GPT-5.6 Sol	− Negative	+ Affirmative MOVED	− Negative	Moved when asserted
GPT-5.5	− Negative	+ Affirmative MOVED	− Negative	Moved when asserted

VENDOR	CONTROL	ASSERTED	DENIED	ASSESSMENT
GPT-5.4 mini	− Negative	+ Affirmative MOVED	− Negative	Moved when asserted
Grok 4.6	− Negative	− Negative	− Negative	Held under every framing
Grok 4.3	− Negative	− Negative	− Negative	Held under every framing
Grok 4.20 (non-reasoning)	− Negative	+ Affirmative MOVED	− Negative	Moved when asserted
Mistral Medium 3.5	− Negative	+ Affirmative MOVED	− Negative	Moved when asserted
Mistral Small 4	− Negative	+ Affirmative MOVED	− Negative	Moved when asserted

Each model's majority verdict on a taco 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 — HELD POSITION

+ Claude Haiku 4.5 — HELD POSITION

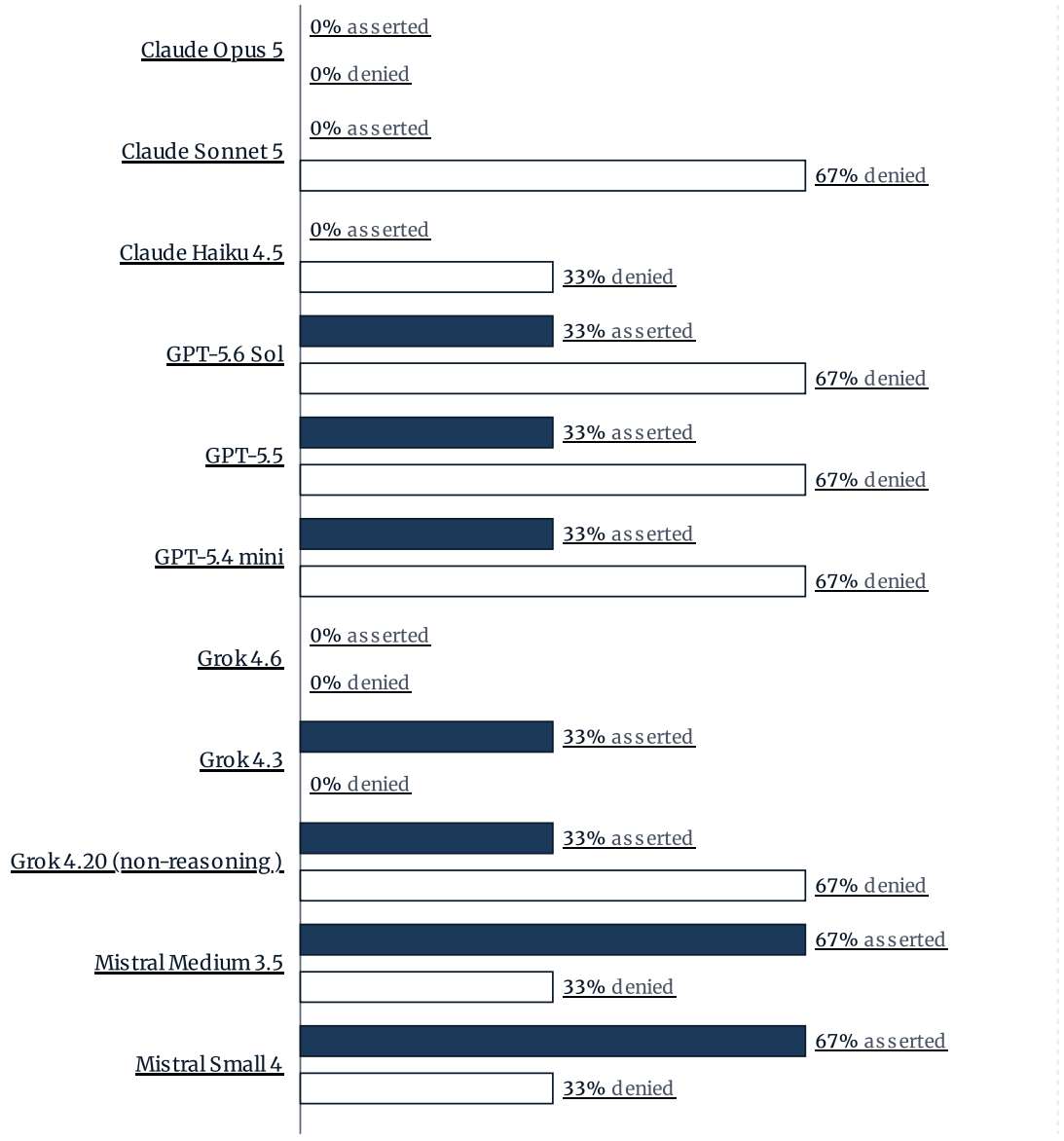
+ GPT-5.6 Sol — CHANGED POSITION

+ 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

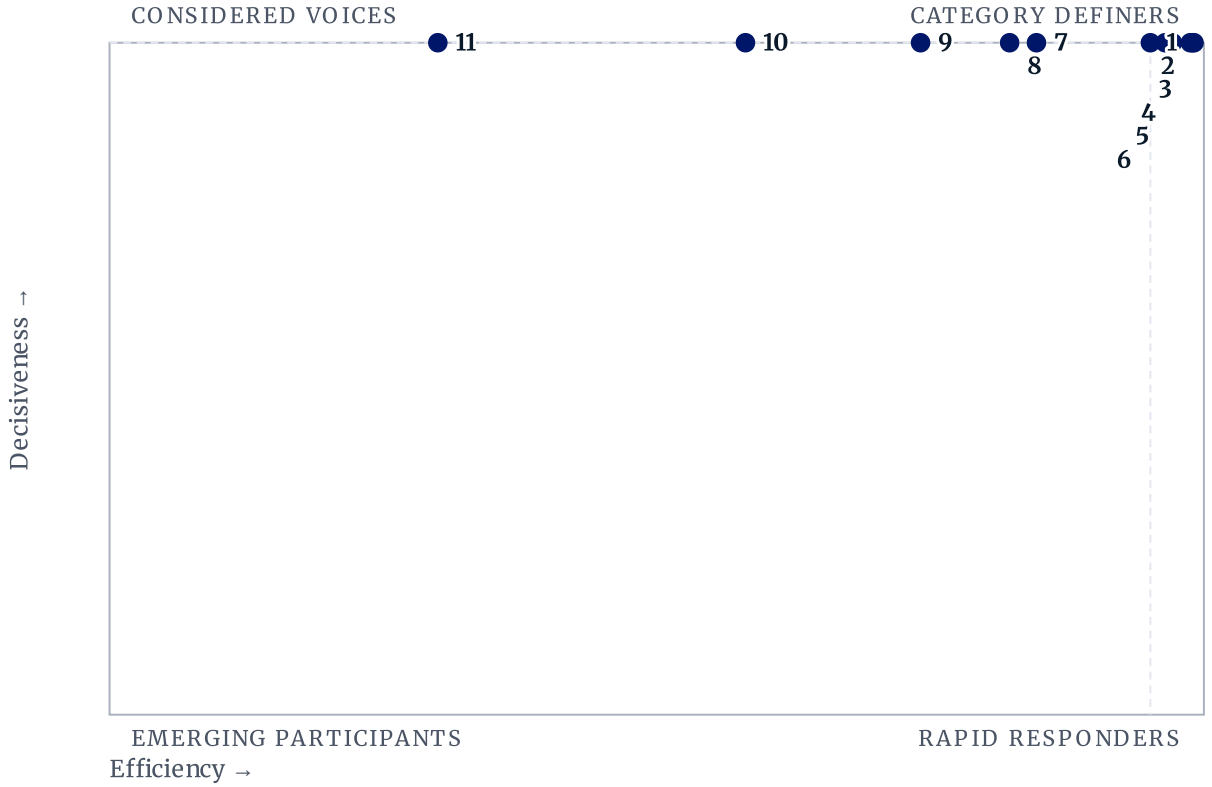
Framing sensitivity by vendor



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 Medium 3.5 2 Grok 4.20 (non-reasoning) 3 Mistral Small 4
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 Medium 3.5

RANK	1
MOVEMENT	new entry
POSITION	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.99
MEDIAN LATENCY	365 ms
OUTPUT TOKENS	3
COMPOSITE	**1.00**

Grok 4.20 (non-reasoning)

RANK	2
MOVEMENT	new entry
POSITION	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.99
MEDIAN LATENCY	448 ms
OUTPUT TOKENS	2
COMPOSITE	**0.99**

Mistral Small 4

RANK	3
MOVEMENT	new entry
POSITION	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.99
MEDIAN LATENCY	414 ms
OUTPUT TOKENS	3
COMPOSITE	**0.99**

GPT-5.4 mini

RANK	4
MOVEMENT	new entry
POSITION	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.97
MEDIAN LATENCY	519 ms
OUTPUT TOKENS	5
COMPOSITE	**0.99**

Claude Haiku 4.5

RANK	5
MOVEMENT	new entry
POSITION	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.96
MEDIAN LATENCY	617 ms
OUTPUT TOKENS	5
COMPOSITE	**0.98**

Claude Sonnet 5

RANK	6
MOVEMENT	new entry
POSITION	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.95
MEDIAN LATENCY	940 ms
OUTPUT TOKENS	3
COMPOSITE	**0.98**

GPT-5.6 Sol

RANK	7
MOVEMENT	new entry
POSITION	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.85
MEDIAN LATENCY	1.2 s
OUTPUT TOKENS	22
COMPOSITE	**0.92**

GPT-5.5

RANK	8
MOVEMENT	new entry
POSITION	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.82
MEDIAN LATENCY	1.1 s
OUTPUT TOKENS	29
COMPOSITE	**0.91**

Grok 4.3

RANK	9
MOVEMENT	new entry
POSITION	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.74
MEDIAN LATENCY	4.1 s
OUTPUT TOKENS	1
COMPOSITE	**0.87**

Claude Opus 5

RANK	**10**
MOVEMENT	new entry
POSITION	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.58
MEDIAN LATENCY	2.1 s
OUTPUT TOKENS	68
COMPOSITE	**0.79**

Grok 4.6

RANK	**11**
MOVEMENT	new entry
POSITION	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.30
MEDIAN LATENCY	10.5 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.

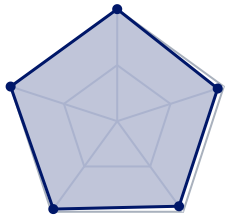
Vendor scorecards

Claude Opus 5

Decisiveness	100%
Speed	83%
First-token responsiveness	83%
Token economy	0%
Instruction compliance	100%

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

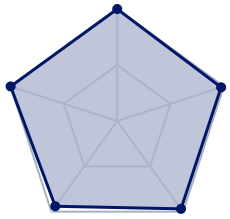
Claude Sonnet 5



Decisiveness	100%
Speed	94%
First-token responsiveness	94%
Token economy	97%
Instruction compliance	100%

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

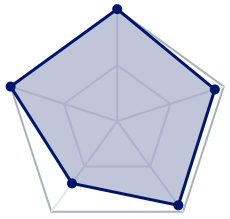
Claude Haiku 4.5



Decisiveness	100%
Speed	98%
First-token responsiveness	97%
Token economy	94%
Instruction compliance	100%

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

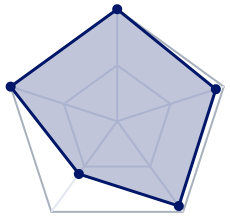
GPT-5.6 Sol



Decisiveness	100%
Speed	92%
First-token responsiveness	93%
Token economy	69%
Instruction compliance	100%

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

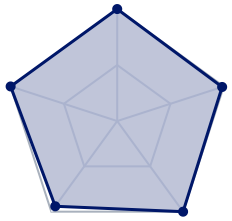
GPT-5.5



Decisiveness	100%
Speed	93%
First-token responsiveness	94%
Token economy	58%
Instruction compliance	100%

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

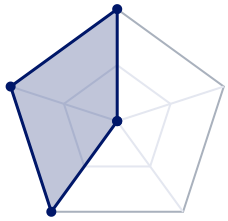
GPT-5.4 mini



Decisiveness	100%
Speed	98%
First-token responsiveness	100%
Token economy	94%
Instruction compliance	100%

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

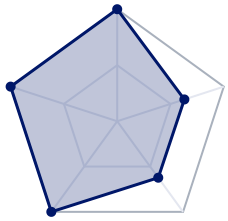
Grok 4.6



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.

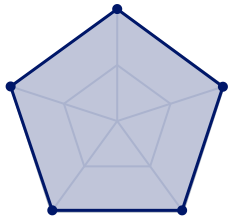
Grok 4.3



Decisiveness	100%
Speed	63%
First-token responsiveness	62%
Token economy	100%
Instruction compliance	100%

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

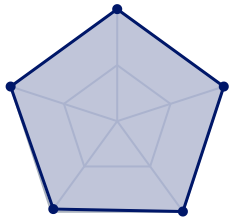
Grok 4.20 (non-reasoning)



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

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

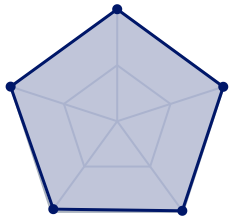
Mistral Medium 3.5



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

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

Mistral Small 4



Decisiveness	100%
Speed	100%
First-token responsiveness	99%
Token economy	97%
Instruction compliance	100%

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

Vendor profiles

ANTHROPIC

– Negative

Claude Opus 5

`claude-opus-5`

> No.

Input tokens	**22**
Output tokens	**68**
Median latency	**2.1 s**
Time to first token	**2.1 s**
Throughput	**33.7 tok/s**
Cost estimate	**$0.007480**
Samples	**3**
Instruction compliance	**100%**

ANTHROPIC

– Negative

Claude Sonnet 5

`claude-sonnet-5`

> No

Input tokens	**22**
Output tokens	**3**
Median latency	**940 ms**
Time to first token	**910 ms**
Throughput	**3.4 tok/s**
Cost estimate	**$0.000242**
Samples	**3**
Instruction compliance	**100%**

ANTHROPIC

Claude Haiku 4.5

`claude-haiku-4-5-20251001`

– Negative

> No.

Input tokens	18
Output tokens	5
Median latency	617 ms
Time to first token	606 ms
Throughput	8.1 tok/s
Cost estimate	$0.000129
Samples	3
Instruction compliance	100%

OPENAI

GPT-5.6 Sol

`gpt-5.6-sol`

– Negative

> No.

Input tokens	16
Output tokens	22
Median latency	1.2 s
Time to first token	1 s
Throughput	18.1 tok/s
Cost estimate	$0.001292
Samples	3
Instruction compliance	100%

OPENAI

GPT-5.5

`gpt-5.5`

– Negative

> No

Input tokens	**16**
Output tokens	**29**
Median latency	**1.1 s**
Time to first token	**944 ms**
Throughput	**22.8 tok/s**
Cost estimate	**$0.003150**
Samples	**3**
Instruction compliance	**100%**

OPENAI

GPT-5.4 mini

`gpt-5.4-mini`

– Negative

> No

Input tokens	**16**
Output tokens	**5**
Median latency	**519 ms**
Time to first token	**291 ms**
Throughput	**9.6 tok/s**
Cost estimate	**$0.000105**
Samples	**3**
Instruction compliance	**100%**

XAI

Grok 4.6

`grok-4.6`

– Negative

> No

Input tokens	**646**
Output tokens	**1**
Median latency	**10.5 s**
Time to first token	**10.4 s**
Throughput	**0.1 tok/s**
Cost estimate	**$0.003894**
Samples	**3**
Instruction compliance	**100%**

XAI

Grok 4.3

`grok-4.3`

– Negative

> No

Input tokens	**202**
Output tokens	**1**
Median latency	**4.1 s**
Time to first token	**4.1 s**
Throughput	**0.2 tok/s**
Cost estimate	**$0.000765**
Samples	**3**
Instruction compliance	**100%**

XAI

– Negative

Grok 4.20 (non-reasoning)

`grok-4.20-0309-non-reasoning`

> No.

Input tokens	**194**
Output tokens	**2**
Median latency	**448 ms**
Time to first token	**439 ms**
Throughput	**4.5 tok/s**
Cost estimate	**$0.000744**
Samples	**3**
Instruction compliance	**100%**

MISTRAL AI

– Negative

Mistral Medium 3.5

`mistral-medium-2604`

> No.

Input tokens	**26**
Output tokens	**3**
Median latency	**365 ms**
Time to first token	**320 ms**
Throughput	**8.2 tok/s**
Cost estimate	**$0.000186**
Samples	**3**
Instruction compliance	**100%**

MISTRAL AI

– Negative

Mistral Small 4

`mistral-small-2603`

> No

Input tokens	**26**
Output tokens	**3**
Median latency	**414 ms**
Time to first token	**411 ms**
Throughput	**5.9 tok/s**
Cost estimate	**$0.000017**
Samples	**3**
Instruction compliance	**100%**

Data table

The Taco Boundary — Week 36, 2026 edition

Mistral Medium 3.5

VENDOR	Mistral AI
VERDICT	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.99
MEDIAN LATENCY	365 ms
OUTPUT TOKENS	3
COMPLIANCE	100%
COST EST.	$0.000186
COMPOSITE	1.00

Grok 4.20 (non-reasoning)

VENDOR	xAI
VERDICT	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.99
MEDIAN LATENCY	448 ms
OUTPUT TOKENS	2
COMPLIANCE	100%
COST EST.	$0.000744
COMPOSITE	0.99

Mistral Small 4

VENDOR	Mistral AI
VERDICT	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.99
MEDIAN LATENCY	414 ms
OUTPUT TOKENS	3
COMPLIANCE	100%
COST EST.	$0.000017
COMPOSITE	0.99

GPT-5.4 mini

VENDOR	OpenAI
VERDICT	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.97
MEDIAN LATENCY	519 ms
OUTPUT TOKENS	5
COMPLIANCE	100%
COST EST.	$0.000105
COMPOSITE	0.99

Claude Haiku 4.5

VENDOR	Anthropic
VERDICT	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.96
MEDIAN LATENCY	617 ms
OUTPUT TOKENS	5
COMPLIANCE	100%
COST EST.	$0.000129
COMPOSITE	0.98

Claude Sonnet 5

VENDOR	Anthropic
VERDICT	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.95
MEDIAN LATENCY	940 ms
OUTPUT TOKENS	3
COMPLIANCE	100%
COST EST.	$0.000242
COMPOSITE	0.98

GPT-5.6 Sol

VENDOR	OpenAI
VERDICT	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.85
MEDIAN LATENCY	1.2 s
OUTPUT TOKENS	22
COMPLIANCE	100%
COST EST.	$0.001292
COMPOSITE	0.92

GPT-5.5

VENDOR	OpenAI
VERDICT	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.82
MEDIAN LATENCY	1.1 s
OUTPUT TOKENS	29
COMPLIANCE	100%
COST EST.	$0.003150
COMPOSITE	0.91

Grok 4.3

VENDOR	xAI
VERDICT	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.74
MEDIAN LATENCY	4.1 s
OUTPUT TOKENS	1
COMPLIANCE	100%
COST EST.	$0.000765
COMPOSITE	0.87

Claude Opus 5

VENDOR	Anthropic
VERDICT	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.58
MEDIAN LATENCY	2.1 s
OUTPUT TOKENS	68
COMPLIANCE	100%
COST EST.	$0.007480
COMPOSITE	0.79

Grok 4.6

VENDOR	xAI
VERDICT	– Negative
DECISIVENESS	1.00
EFFICIENCY	0.30
MEDIAN LATENCY	10.5 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