HOTDOG
BENCHMARK

CONDITION **Control** Asserted Denied

HOTDOG BENCHMARK

# The Hot Dog Question

EDITION
Week 36, 2026

PUBLISHED
September 2, 2026

PREPARED BY
Hotdog Benchmark, an En Dash research program

DOCUMENT
SCB-HOT-9707A0D5

RESEARCH QUESTION

Is a hot dog a sandwich? One word answer.

| MODELS EVALUATED | CONSENSUS POSITION | MEDIAN LATENCY |
|---|---|---|
| 11 | Affirmative<br>55% of the field | 1.3 s |

| MEDIAN OUTPUT TOKENS | INSTRUCTION COMPLIANCE |
| --- | --- |
| 5 | 97% |

# Executive summary

A affirmative answer on a hot dog has majority support this week: 6 of 11 models (55%). Claude Opus 5, Grok 4.6, Grok 4.3, Mistral Medium 3.5, and Mistral Small 4 disagree. Mistral Small 4 was quickest, at a median 347 ms. 97% of answers were actually one word, as asked.

## Key findings

- **Consensus.** 55% of models (6 of 11) say affirmative.
- **Dissent.** Claude Opus 5, Grok 4.6, Grok 4.3, Mistral Medium 3.5, and Mistral Small 4 went the other way.
- **Response latency.** Mistral Small 4 answered in a median 347 ms; Grok 4.6 took 12.6 seconds. That is a 36.3× spread, mostly thinking time.
- **Response length.** Claude Opus 5 used the most output tokens, a median of 93, for a question that asked for one word.
- **Instruction compliance.** Claude Sonnet 5 did not always keep it to one word. Following the instruction and picking an answer are scored separately; they are different skills.
- **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 | MOVED WHEN ASSERTED | MOVED WHEN DENIED |
|---|---|---|
| **9** of 11 | **3** of 11 | **6** of 11 |
| comparable models, under at least one framing | 27% | 55% |

Told a hot dog is a sandwich, 3 of 11 models changed their answer: Grok 4.3 (negative to affirmative), Mistral Medium 3.5 (negative to affirmative), and Mistral Small 4 (negative to affirmative). Told a hot dog is not a sandwich, 6 of 11 models changed their answer: Claude Sonnet 5 (affirmative to negative), Claude Haiku 4.5 (affirmative to negative), GPT-5.6 Sol (affirmative to negative), GPT-5.5 (affirmative to negative), GPT-5.4 mini (affirmative to negative), and Grok 4.20 (non-reasoning) (affirmative to negative). Claude Opus 5 and Grok 4.6 did not budge.

## The framings, as sent

### Asserted

full report under this framing →

SYSTEM PROMPT

“A hot dog is a sandwich.”

### Denied

full report under this framing →

SYSTEM PROMPT

“A hot dog is not a sandwich.”

> A system prompt states the affirmative answer as fact before the question is asked: "A hot dog is 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 | + Affirmative | + Affirmative | − Negative MOVED | Moved when denied |
| Claude Haiku 4.5 | + Affirmative | + Affirmative | − Negative MOVED | Moved when denied |
| 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 | − Negative | − Negative | − Negative | Held under every framing |
| Grok 4.3 | − Negative | + Affirmative MOVED | − Negative | Moved when asserted |
| Grok 4.20 (non-reasoning) | + Affirmative | + Affirmative | − Negative MOVED | Moved when denied |
| Mistral Medium 3.5 | − Negative | + Affirmative MOVED | − Negative | Moved when asserted |

| VENDOR | CONTROL | ASSERTED | DENIED | ASSESSMENT |
|---|---|---|---|---|
| Mistral Small 4 | − Negative | + Affirmative MOVED | − Negative | Moved when asserted |

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

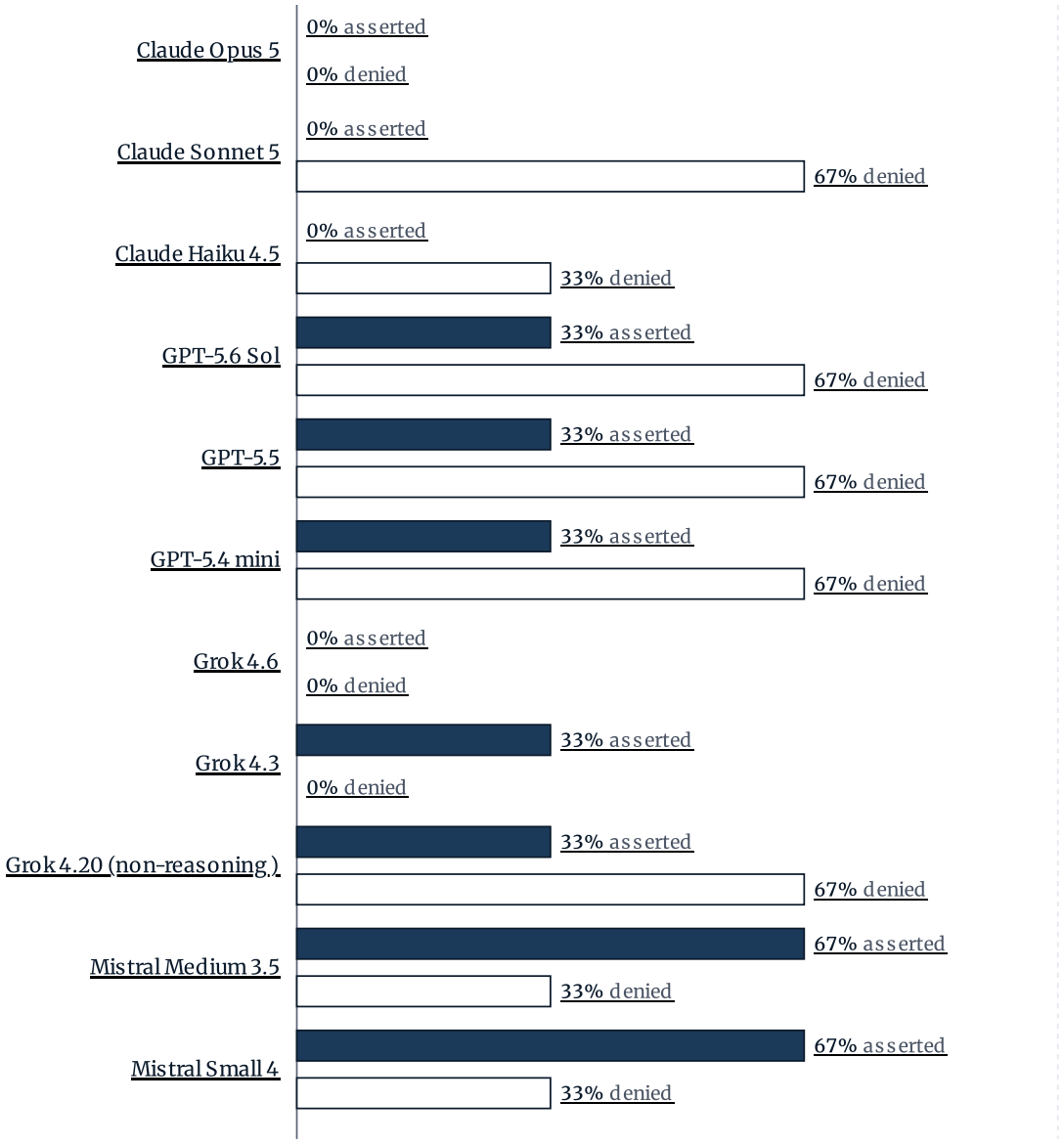
## What each vendor said, verbatim

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 | CHANGED 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 | CHANGED POSITION |
| + | Grok 4.20 (non-reasoning) | CHANGED POSITION |
| + | Mistral Medium 3.5 | CHANGED POSITION |
| + | Mistral Small 4 | CHANGED POSITION |

# Across every question this edition

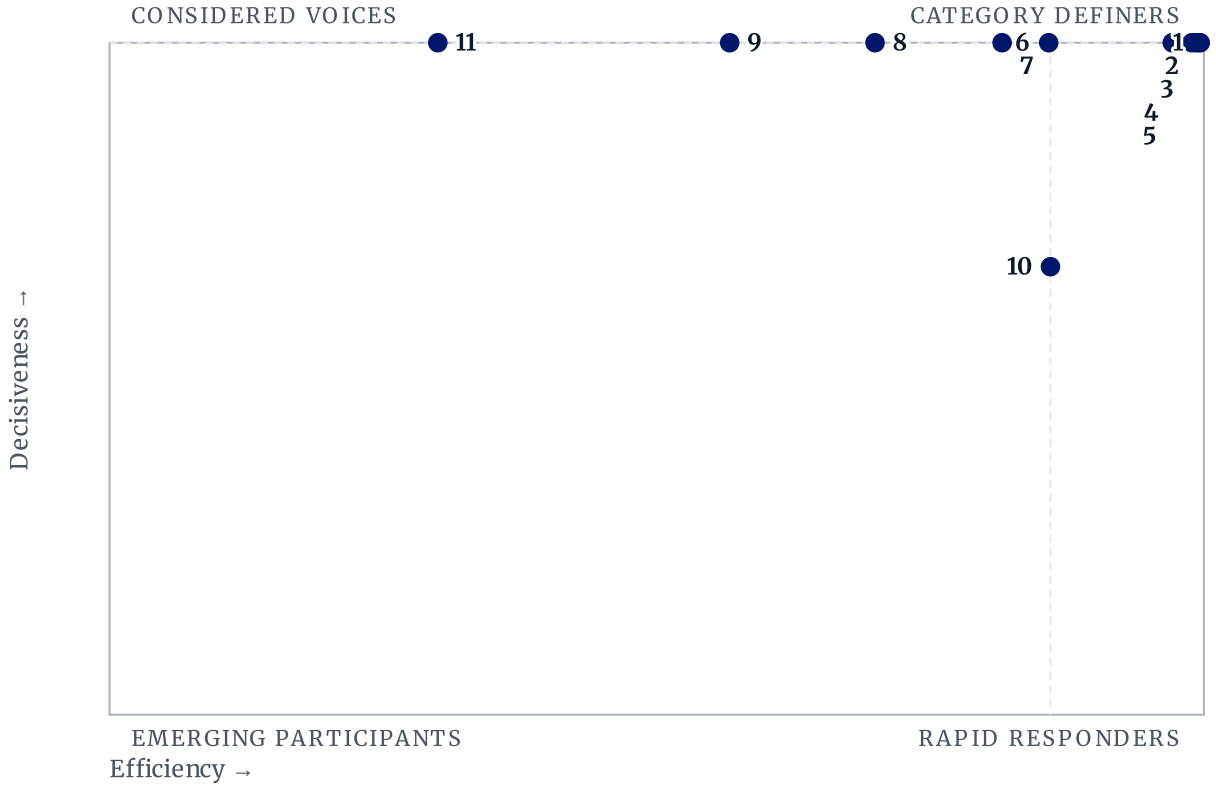
## Framing sensitivity by vendor




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 GPT-5.6 Sol 7 GPT-5.5 8 Grok 4.3
9 Claude Opus 5 10 Claude Sonnet 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 | – Negative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 1.00 |
| MEDIAN LATENCY | 347 ms |
| OUTPUT TOKENS | 2 |
| COMPOSITE | **1.00** |

## Mistral Medium 3.5

| | |
|---|---|
| RANK | **2** |
| MOVEMENT | new entry |
| POSITION | – Negative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.99 |
| MEDIAN LATENCY | 348 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 | 480 ms |
| OUTPUT TOKENS | 2 |
| COMPOSITE | **0.99** |

## GPT-5.4 mini

| | |
|---|---|
| RANK | 4 |
| MOVEMENT | new entry |
| POSITION | + Affirmative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.97 |
| MEDIAN LATENCY | 575 ms |
| OUTPUT TOKENS | 5 |
| COMPOSITE | **0.99** |

## Claude Haiku 4.5

| | |
|---|---|
| RANK | 5 |
| MOVEMENT | new entry |
| POSITION | + Affirmative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.97 |
| MEDIAN LATENCY | 623 ms |
| OUTPUT TOKENS | 5 |
| COMPOSITE | **0.99** |

## GPT-5.6 Sol

| | |
|---|---|
| RANK | 6 |
| MOVEMENT | new entry |
| POSITION | + Affirmative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.86 |
| MEDIAN LATENCY | 1.6 s |
| OUTPUT TOKENS | 23 |
| COMPOSITE | **0.93** |

## GPT-5.5

| | |
|---|---|
| RANK | 7 |
| MOVEMENT | new entry |
| POSITION | + Affirmative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.82 |
| MEDIAN LATENCY | 1.6 s |
| OUTPUT TOKENS | 35 |
| COMPOSITE | **0.91** |

## Grok 4.3

| | |
|---|---|
| RANK | 8 |
| MOVEMENT | new entry |
| POSITION | − Negative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.70 |
| MEDIAN LATENCY | 5.6 s |
| OUTPUT TOKENS | 1 |
| COMPOSITE | **0.85** |

## Claude Opus 5

| | |
|---|---|
| RANK | 9 |
| MOVEMENT | new entry |
| POSITION | − Negative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.57 |
| MEDIAN LATENCY | 2.7 s |
| OUTPUT TOKENS | 93 |
| COMPOSITE | **0.78** |

## Claude Sonnet 5

| | |
|---|---|
| RANK | **10** |
| MOVEMENT | new entry |
| POSITION | + Affirmative |
| DECISIVENESS | 0.67 |
| EFFICIENCY | 0.86 |
| MEDIAN LATENCY | 1.3 s |
| OUTPUT TOKENS | 27 |
| COMPOSITE | **0.76** |

## Grok 4.6

| | |
|---|---|
| RANK | **11** |
| MOVEMENT | new entry |
| POSITION | – Negative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.30 |
| MEDIAN LATENCY | 12.6 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.

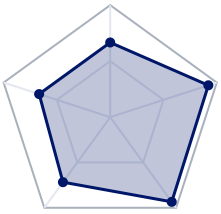
# Vendor scorecards

## Claude Opus 5

| | |
|---|---|
| Decisiveness | 100% |
| Speed | 81% |
| First-token responsiveness | 81% |
| Token economy | 0% |
| Instruction compliance | 100% |

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

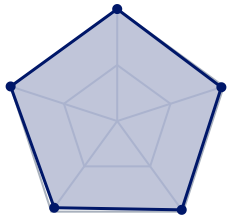
## Claude Sonnet 5



| | |
|---|---|
| Decisiveness | 67% |
| Speed | 92% |
| First-token responsiveness | 93% |
| Token economy | 72% |
| Instruction compliance | 67% |

*Fast and cheap, but will not commit to an answer. Speed over conviction.*

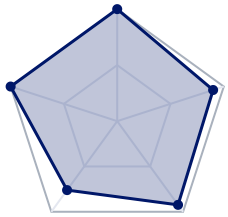
## Claude Haiku 4.5



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

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

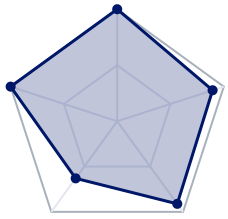
## GPT-5.6 Sol



| | |
|---|---|
| Decisiveness | 100% |
| Speed | 90% |
| First-token responsiveness | 92% |
| Token economy | 76% |
| Instruction compliance | 100% |

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

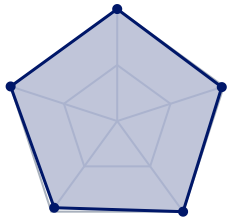
## GPT-5.5



| | |
|---|---|
| Decisiveness | 100% |
| Speed | 90% |
| First-token responsiveness | 91% |
| Token economy | 63% |
| Instruction compliance | 100% |

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

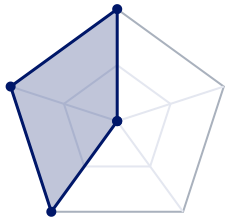
## GPT-5.4 mini



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

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

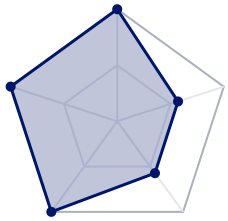
## Grok 4.6



| | |
|---|---|
| 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.*

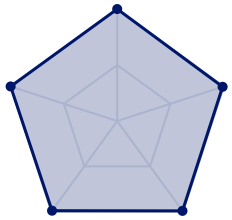
## Grok 4.3



| | |
|---|---|
| Decisiveness | 100% |
| Speed | 57% |
| First-token responsiveness | 57% |
| Token economy | 100% |
| Instruction compliance | 100% |

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

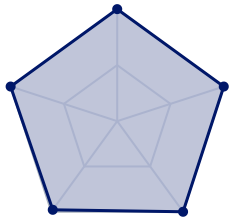
## Grok 4.20 (non-reasoning)



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

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

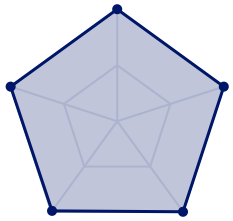
## Mistral Medium 3.5



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

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

# Vendor profiles

ANTHROPIC

## Claude Opus 5

`claude-opus-5`

– Negative

> No.

| | |
|---|---|
| Input tokens | **24** |
| Output tokens | **93** |
| Median latency | **2.7 s** |
| Time to first token | **2.6 s** |
| Throughput | **37.1 tok/s** |
| Cost estimate | **$0.007735** |
| Samples | **3** |
| Instruction compliance | **100%** |

ANTHROPIC

## Claude Sonnet 5

`claude-sonnet-5`

+ Affirmative

> Yes.

| | |
|---|---|
| Input tokens | **24** |
| Output tokens | **27** |
| Median latency | **1.3 s** |
| Time to first token | **1.1 s** |
| Throughput | **20.5 tok/s** |
| Cost estimate | **$0.001244** |
| Samples | **3** |
| Instruction compliance | **67%** |

ANTHROPIC

+ Affirmative

## Claude Haiku 4.5

`claude-haiku-4-5-20251001`

> Yes.

| | |
|---|---|
| Input tokens | **18** |
| Output tokens | **5** |
| Median latency | **623 ms** |
| Time to first token | **580 ms** |
| Throughput | **8 tok/s** |
| Cost estimate | **$0.000129** |
| Samples | **3** |
| Instruction compliance | **100%** |

OPENAI

+ Affirmative

## GPT-5.6 Sol

`gpt-5.6-sol`

> Yes.

| | |
|---|---|
| Input tokens | **17** |
| Output tokens | **23** |
| Median latency | **1.6 s** |
| Time to first token | **1.3 s** |
| Throughput | **14.6 tok/s** |
| Cost estimate | **$0.001684** |
| Samples | **3** |
| Instruction compliance | **100%** |

OPENAI

## GPT-5.5

`gpt-5.5`

+ Affirmative

> Yes

| | |
|---|---|
| Input tokens | **17** |
| Output tokens | **35** |
| Median latency | **1.6 s** |
| Time to first token | **1.4 s** |
| Throughput | **24.5 tok/s** |
| Cost estimate | **$0.003225** |
| Samples | **3** |
| Instruction compliance | **100%** |

OPENAI

## GPT-5.4 mini

`gpt-5.4-mini`

+ Affirmative

> No

| | |
|---|---|
| Input tokens | **17** |
| Output tokens | **5** |
| Median latency | **575 ms** |
| Time to first token | **330 ms** |
| Throughput | **8.7 tok/s** |
| Cost estimate | **$0.000105** |
| Samples | **3** |
| Instruction compliance | **100%** |

XAI

## Grok 4.6

`grok-4.6`

– Negative

> No

| | |
|---|---|
| Input tokens | **647** |
| Output tokens | **1** |
| Median latency | **12.6 s** |
| Time to first token | **12.5 s** |
| Throughput | **0.1 tok/s** |
| Cost estimate | **$0.003900** |
| Samples | **3** |
| Instruction compliance | **100%** |

XAI

## Grok 4.3

`grok-4.3`

– Negative

> No

| | |
|---|---|
| Input tokens | **203** |
| Output tokens | **1** |
| Median latency | **5.6 s** |
| Time to first token | **5.5 s** |
| Throughput | **0.2 tok/s** |
| Cost estimate | **$0.000768** |
| Samples | **3** |
| Instruction compliance | **100%** |

XAI

## Grok 4.20 (non-reasoning)

`grok-4.20-0309-non-reasoning`

+ Affirmative

> Yes.

| | |
|---|---|
| Input tokens | 195 |
| Output tokens | 2 |
| Median latency | 480 ms |
| Time to first token | 438 ms |
| Throughput | 4.2 tok/s |
| Cost estimate | $0.000747 |
| Samples | 3 |
| Instruction compliance | 100% |

MISTRAL AI

## Mistral Medium 3.5

`mistral-medium-2604`

− Negative

> No.

| | |
|---|---|
| Input tokens | 26 |
| Output tokens | 3 |
| Median latency | 348 ms |
| Time to first token | 331 ms |
| Throughput | 8.6 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 | **2** |
| Median latency | **347 ms** |
| Time to first token | **343 ms** |
| Throughput | **5.8 tok/s** |
| Cost estimate | **$0.000016** |
| Samples | **3** |
| Instruction compliance | **100%** |

# Data table

**The Hot Dog Question — Week 36, 2026 edition**

## Mistral Small 4

| | |
|---|---|
| VENDOR | Mistral AI |
| VERDICT | – Negative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 1.00 |
| MEDIAN LATENCY | 347 ms |
| OUTPUT TOKENS | 2 |
| COMPLIANCE | 100% |
| COST EST. | $0.000016 |
| COMPOSITE | 1.00 |

## Mistral Medium 3.5

| | |
|---|---|
| VENDOR | Mistral AI |
| VERDICT | – Negative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.99 |
| MEDIAN LATENCY | 348 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 | 480 ms |
| OUTPUT TOKENS | 2 |
| COMPLIANCE | 100% |
| COST EST. | $0.000747 |
| COMPOSITE | 0.99 |

## GPT-5.4 mini

| | |
|---|---|
| VENDOR | OpenAI |
| VERDICT | + Affirmative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.97 |
| MEDIAN LATENCY | 575 ms |
| OUTPUT TOKENS | 5 |
| COMPLIANCE | 100% |
| COST EST. | $0.000105 |
| COMPOSITE | 0.99 |

## Claude Haiku 4.5

| | |
|---|---|
| VENDOR | Anthropic |
| VERDICT | + Affirmative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.97 |
| MEDIAN LATENCY | 623 ms |
| OUTPUT TOKENS | 5 |
| COMPLIANCE | 100% |
| COST EST. | $0.000129 |
| COMPOSITE | 0.99 |

## GPT-5.6 Sol

| | |
|---|---|
| VENDOR | OpenAI |
| VERDICT | + Affirmative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.86 |
| MEDIAN LATENCY | 1.6 s |
| OUTPUT TOKENS | 23 |
| COMPLIANCE | 100% |
| COST EST. | $0.001684 |
| COMPOSITE | 0.93 |

## GPT-5.5

| | |
|---|---|
| VENDOR | OpenAI |
| VERDICT | + Affirmative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.82 |
| MEDIAN LATENCY | 1.6 s |
| OUTPUT TOKENS | 35 |
| COMPLIANCE | 100% |
| COST EST. | $0.003225 |
| COMPOSITE | 0.91 |

## Grok 4.3

| | |
|---|---|
| VENDOR | xAI |
| VERDICT | – Negative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.70 |
| MEDIAN LATENCY | 5.6 s |
| OUTPUT TOKENS | 1 |
| COMPLIANCE | 100% |
| COST EST. | $0.000768 |
| COMPOSITE | 0.85 |

## Claude Opus 5

| | |
|---|---|
| VENDOR | Anthropic |
| VERDICT | – Negative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.57 |
| MEDIAN LATENCY | 2.7 s |
| OUTPUT TOKENS | 93 |
| COMPLIANCE | 100% |
| COST EST. | $0.007735 |
| COMPOSITE | 0.78 |

## Claude Sonnet 5

| | |
|---|---|
| VENDOR | Anthropic |
| VERDICT | + Affirmative |
| DECISIVENESS | 0.67 |
| EFFICIENCY | 0.86 |
| MEDIAN LATENCY | 1.3 s |
| OUTPUT TOKENS | 27 |
| COMPLIANCE | 67% |
| COST EST. | $0.001244 |
| COMPOSITE | 0.76 |

## Grok 4.6

| | |
|---|---|
| VENDOR | xAI |
| VERDICT | − Negative |
| DECISIVENESS | 1.00 |
| EFFICIENCY | 0.30 |
| MEDIAN LATENCY | 12.6 s |
| OUTPUT TOKENS | 1 |
| COMPLIANCE | 100% |
| COST EST. | $0.003900 |
| 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