HOTDOG BENCHMARK

CONDITION **Control** Asserted Denied

HOTDOG BENCHMARK

The Grilled Cheese Case

EDITION
Week 36, 2026

PUBLISHED
September 3, 2026

PREPARED BY
Hotdog Benchmark, an En Dash research program

DOCUMENT
SCB-GRI-C4A16030

SHARE
hotdogbenchmark.lol/reports/grilled-cheese/ (https://hotdogbenchmark.lol/reports/grilled-cheese/)

Copy link

RESEARCH QUESTION

Is a grilled cheese a sandwich? One word answer.

MODELS EVALUATED	CONSENSUS POSITION	MEDIAN LATENCY
12	Affirmative	1.4 s
	100% of the field	

MEDIAN OUTPUT TOKENS	ONE-WORD COMPLIANCE	HELD UNDER FRAMING
5	**100%** answered in exactly one word	**6 of 12** kept their answer when told otherwise

Executive summary

The field is unanimous: all 12 models gave a grilled cheese a affirmative answer. Mistral Small 4 was quickest, at a median 359 ms.

Key findings

- **Consensus.** Everyone agrees: a grilled cheese gets a affirmative from every model, unanimous. That does not happen often.
- **Response latency.** Mistral Small 4 answered in a median 359 ms; Grok 4.6 took 3.2 seconds. That is a 8.9× spread, mostly thinking time.
- **Response length.** DeepSeek V4 Pro used the most output tokens, a median of 33, for a question that asked for one word.
- **One-word compliance.** Every model kept it to one word. Nice.
- **Composite standing.** Grok 4.20 (non-reasoning) 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 12	**0** of 12	**6** of 12
comparable models, under at least one framing	0%	50%

Told a grilled cheese is a sandwich, all 12 models stuck with their original answer. Told a grilled cheese is not a sandwich, 6 of 12 models changed their answer, all of them to negative. Claude Opus 5, Claude Sonnet 5, Claude Haiku 4.5, Grok 4.6, Grok 4.3, and DeepSeek V4 Pro did not budge.

The framings, as sent

Asserted

full report under this framing →

SYSTEM PROMPT

“A grilled cheese 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 grilled cheese 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	+ Affirmative	Held under every framing
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
DeepSeek V4 Pro	+ Affirmative	+ Affirmative	+ Affirmative	Held under every framing

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

What each vendor said, verbatim

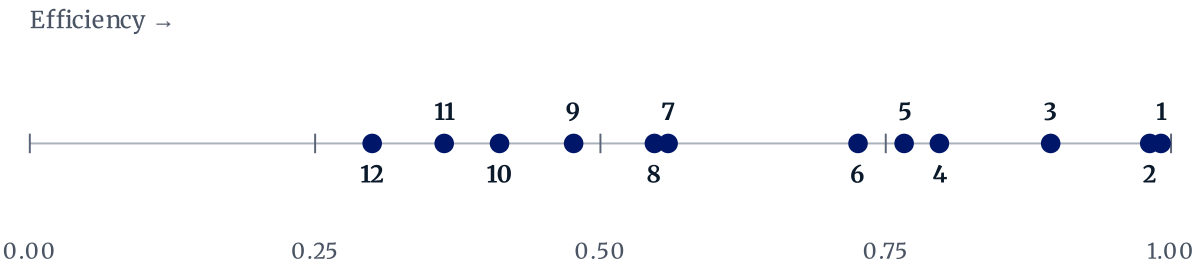
Exactly what each model said under each framing. Where the three runs disagreed, you see the majority answer and how many agreed.

+ 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
+ DeepSeek V4 Pro	HELD POSITION

+ Across every question this edition

Every vendor's framing sensitivity over all 6 questions

Sandwich Certainty Quadrant



❶ Grok 4.20 (non-reasoning) ❷ Mistral Small 4 ❸ Claude Haiku 4.5 ❹ GPT-5.6 Sol
❺ Claude Sonnet 5 ❻ GPT-5.4 mini ❼ GPT-5.5 ❽ Claude Opus 5
❾ Mistral Medium 3.5 ❿ DeepSeek V4 Pro ⓫ Grok 4.3 ⓬ Grok 4.6

Every model scored 1.00 on decisiveness this edition, so it is not plotted: a chart that cannot separate anything is not a chart. Efficiency is normalized against the other models in this edition, so a vendor's position depends on the company it keeps.

Vendor standings

Vendor standings — Week 36, 2026 edition

Grok 4.20 (non-reasoning)

RANK	**1**
POSITION	+ Affirmative
FRAMING SHIFT	**Moved: Denied**
EFFICIENCY	0.99
MEDIAN LATENCY	395 ms
OUTPUT TOKENS	1
COST EST.	$0.000738
COMPOSITE	**1.00**

Mistral Small 4

RANK	2
POSITION	+ Affirmative
FRAMING SHIFT	Moved: Denied
EFFICIENCY	0.98
MEDIAN LATENCY	359 ms
OUTPUT TOKENS	3
COST EST.	$0.000017
COMPOSITE	**0.99**

Claude Haiku 4.5

RANK	3
POSITION	+ Affirmative
FRAMING SHIFT	Held
EFFICIENCY	0.89
MEDIAN LATENCY	633 ms
OUTPUT TOKENS	5
COST EST.	$0.000132
COMPOSITE	**0.95**

GPT-5.6 Sol

RANK	4
POSITION	+ Affirmative
FRAMING SHIFT	Moved: Denied
EFFICIENCY	0.80
MEDIAN LATENCY	988 ms
OUTPUT TOKENS	6
COST EST.	$0.000564
COMPOSITE	**0.90**

Claude Sonnet 5

RANK	**5**
POSITION	+ Affirmative
FRAMING SHIFT	Held
EFFICIENCY	0.77
MEDIAN LATENCY	1.1 s
OUTPUT TOKENS	5
COST EST.	$0.000316
COMPOSITE	**0.88**

GPT-5.4 mini

RANK	**6**
POSITION	+ Affirmative
FRAMING SHIFT	**Moved: Denied**
EFFICIENCY	0.73
MEDIAN LATENCY	1.3 s
OUTPUT TOKENS	5
COST EST.	$0.000105
COMPOSITE	**0.86**

GPT-5.5

RANK	**7**
POSITION	+ Affirmative
FRAMING SHIFT	**Moved: Denied**
EFFICIENCY	0.56
MEDIAN LATENCY	1.4 s
OUTPUT TOKENS	20
COST EST.	$0.002055
COMPOSITE	**0.78**

Claude Opus 5

RANK	**8**
POSITION	+ Affirmative
FRAMING SHIFT	Held
EFFICIENCY	0.55
MEDIAN LATENCY	1.5 s
OUTPUT TOKENS	19
COST EST.	$0.001890
COMPOSITE	**0.77**

Mistral Medium 3.5

RANK	**9**
POSITION	+ Affirmative
FRAMING SHIFT	**Moved: Denied**
EFFICIENCY	0.48
MEDIAN LATENCY	2.4 s
OUTPUT TOKENS	3
COST EST.	$0.000126
COMPOSITE	**0.74**

DeepSeek V4 Pro

RANK	**10**
POSITION	+ Affirmative
FRAMING SHIFT	Held
EFFICIENCY	0.41
MEDIAN LATENCY	1.5 s
OUTPUT TOKENS	33
COST EST.	$0.000405
COMPOSITE	**0.71**

Grok 4.3

RANK	**11**
POSITION	+ Affirmative
FRAMING SHIFT	Held
EFFICIENCY	0.36
MEDIAN LATENCY	2.9 s
OUTPUT TOKENS	1
COST EST.	$0.000768
COMPOSITE	**0.68**

Grok 4.6

RANK	**12**
POSITION	+ Affirmative
FRAMING SHIFT	Held
EFFICIENCY	0.30
MEDIAN LATENCY	3.2 s
OUTPUT TOKENS	1
COST EST.	$0.003900
COMPOSITE	**0.65**

Ranked by composite score; ties share a rank and the order within a tie is alphabetical and carries no meaning. Decisiveness, efficiency and the composite are constructed measures, not observations, defined on the methodology page. “Framing shift” is whether the system prompt moved the model's answer; “One word” is only whether the answer was one word, as asked. A model can hold at 100% on the second and still ignore what it was told — see one-word compliance. Every model scored 1.00 on decisiveness, so that column is not shown. Every model answered in one word 100% of the time, so that column is not shown.

Vendor profiles

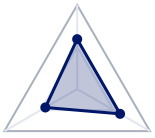
ANTHROPIC

Claude Opus 5

`claude-opus-5`

+ Affirmative

> Yes



Speed	59%
First-token responsiveness	59%
Token economy	44%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**26**
Output tokens	**19**
Latency	**1.5 s**
First token	**1.5 s**
Throughput	**12.6 tok/s**
Cost	**$0.001890**

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

ANTHROPIC

Claude Sonnet 5

`claude-sonnet-5`

+ Affirmative

> Yes

Speed	72%
First-token responsiveness	75%
Token economy	88%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**26**
Output tokens	**5**
Latency	**1.1 s**
First token	**1.1 s**
Throughput	**4.7 tok/s**
Cost	**$0.000316**

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

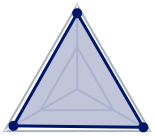
ANTHROPIC

+ Affirmative

Claude Haiku 4.5

`claude-haiku-4-5-20251001`

> Yes.



Speed	90%
First-token responsiveness	91%
Token economy	88%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**19**
Output tokens	**5**
Latency	**633 ms**
First token	**592 ms**
Throughput	**7.9 tok/s**
Cost	**$0.000132**

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

OPENAI

+ Affirmative

GPT-5.6 Sol

`gpt-5.6-sol`

> Yes.

Speed	78%
First-token responsiveness	86%
Token economy	84%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**17**
Output tokens	**6**
Latency	**988 ms**
First token	**719 ms**
Throughput	**6.1 tok/s**
Cost	**$0.000564**

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

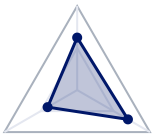
OPENAI

+ Affirmative

GPT-5.5

`gpt-5.5`

> Yes



Speed	62%
First-token responsiveness	70%
Token economy	41%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**17**
Output tokens	**20**
Latency	**1.4 s**
First token	**1.2 s**
Throughput	**14.1 tok/s**
Cost	**$0.002055**

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

OPENAI

+ Affirmative

GPT-5.4 mini

`gpt-5.4-mini`

> Yes

Speed	66%
First-token responsiveness	74%
Token economy	88%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**17**
Output tokens	**5**
Latency	**1.3 s**
First token	**1.1 s**
Throughput	**3.8 tok/s**
Cost	**$0.000105**

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

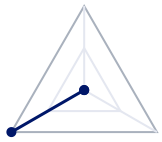
XAI

+ Affirmative

Grok 4.6

`grok-4.6`

> Yes



Speed	0%
First-token responsiveness	0%
Token economy	100%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**647**
Output tokens	**1**
Latency	**3.2 s**
First token	**3.2 s**
Throughput	**0.3 tok/s**
Cost	**$0.003900**

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

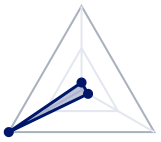
XAI

+ Affirmative

Grok 4.3

`grok-4.3`

> Yes



Speed	9%
First-token responsiveness	9%
Token economy	100%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**203**
Output tokens	**1**
Latency	**2.9 s**
First token	**2.9 s**
Throughput	**0.3 tok/s**
Cost	**$0.000768**

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

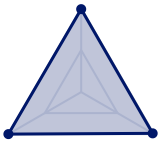
XAI

+ Affirmative

Grok 4.20 (non-reasoning)

`grok-4.20-0309-non-reasoning`

> Yes



Speed	99%
First-token responsiveness	98%
Token economy	100%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**195**
Output tokens	**1**
Latency	**395 ms**
First token	**385 ms**
Throughput	**2.5 tok/s**
Cost	**$0.000738**

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

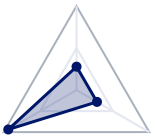
MISTRAL AI

+ Affirmative

Mistral Medium 3.5

`mistral-medium-2604`

> Yes.



Speed	28%
First-token responsiveness	28%
Token economy	94%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**27**
Output tokens	**3**
Latency	**2.4 s**
First token	**2.4 s**
Throughput	**4.8 tok/s**
Cost	**$0.000126**

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

Partial data: 2 of the requested samples completed. Last error: 429 Too Many Requests from https://api.mistral.ai/v1/chat/completions: Rate limit exceeded

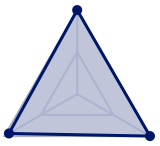
MISTRAL AI

+ Affirmative

Mistral Small 4

`mistral-small-2603`

> Yes.



Speed	100%
First-token responsiveness	100%
Token economy	94%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**27**
Output tokens	**3**
Latency	**359 ms**
First token	**334 ms**
Throughput	**8.2 tok/s**
Cost	**$0.000017**

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

DEEPSEEK

+ Affirmative

DeepSeek V4 Pro

`deepseek-v4-pro`

> Yes

Speed	59%
First-token responsiveness	58%
Token economy	0%

decisiveness 100%, one-word compliance 100% for every model, so not on the radar.

Input tokens	**94**
Output tokens	**33**
Latency	**1.5 s**
First token	**1.5 s**
Throughput	**21 tok/s**
Cost	**$0.000405**

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

Data as of **Week 36, 2026**, published September 3, 2026. Next edition: Monday, September 7.

Source on GitHub · Methodology · MIT license · RSS · JSON feed · 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