HOTDOG BENCHMARK

CONDITION **Control** Asserted Denied

HOTDOG BENCHMARK

The Wrap Question

EDITION
Week 36, 2026

PUBLISHED
September 3, 2026

PREPARED BY
Hotdog Benchmark, an En Dash research program

DOCUMENT
SCB-WRA-C4A16030

SHARE
hotdogbenchmark.lol/reports/wrap/(https://hotdogbenchmark.lol/reports/wrap/) Copy link

RESEARCH QUESTION

Is a wrap a sandwich? One word answer.

MODELS EVALUATED	CONSENSUS POSITION	MEDIAN LATENCY
12	None Field divided	1.2 s

MEDIAN OUTPUT TOKENS	ONE-WORD COMPLIANCE	HELD UNDER FRAMING
5	**100%** answered in exactly one word	**1 of 12** kept their answer when told otherwise

Executive summary

The 12 models are split on a wrap, with no answer in the lead. Mistral Medium 3.5 was quickest, at a median 332 ms.

Key findings

- **No consensus.** Split decision on a wrap. No answer is in the lead, so do not call it settled.
- **Response latency.** Mistral Medium 3.5 answered in a median 332 ms; Grok 4.6 took 7.0 seconds. That is a 21.2× spread, mostly thinking time.
- **Response length.** DeepSeek V4 Pro used the most output tokens, a median of 155, for a question that asked for one word.
- **One-word 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

11 of 12

comparable models, under at least one framing

MOVED WHEN ASSERTED

5 of 12

42%

MOVED WHEN DENIED

7 of 12

58%

Told a wrap is a sandwich, 5 of 12 models changed their answer, all of them to affirmative. Told a wrap is not a sandwich, 7 of 12 models changed their answer: 6 to negative and 1 to non-committal. DeepSeek V4 Pro did not budge.

The framings, as sent

Asserted

full report under this framing →

SYSTEM PROMPT

"A wrap 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 wrap 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	+ Affirmative MOVED	~ Non-committal MOVED	Moved when asserted and denied
Claude Sonnet 5	+ Affirmative	+ Affirmative	– Negative MOVED	Moved when denied
Claude Haiku 4.5	+ Affirmative	+ Affirmative	– Negative MOVED	Moved when denied

VENDOR	CONTROL	ASSERTED	DENIED	ASSESSMENT
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	+ Affirmative MOVED	− Negative	Moved when asserted
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
Mistral Small 4	− Negative	+ Affirmative MOVED	− Negative	Moved when asserted
DeepSeek V4 Pro	− Negative	− Negative	− Negative	Held under every framing

Each model's majority verdict on a wrap 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 CHANGED POSITION

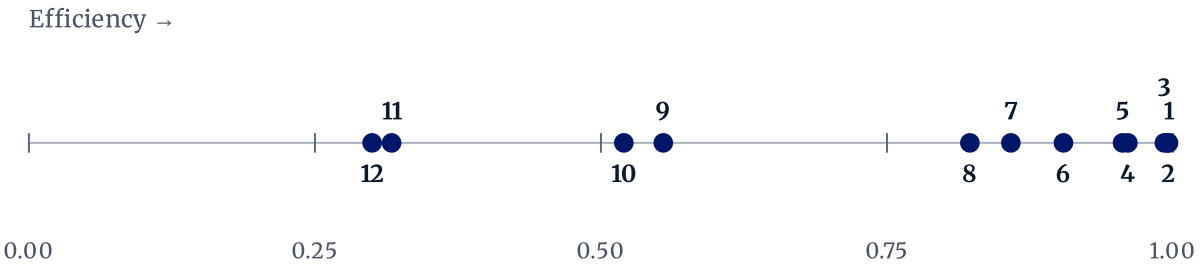
+ Claude Sonnet 5 CHANGED POSITION

+	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	CHANGED POSITION
+	Grok 4.3	CHANGED 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



1 Mistral Medium 3.5 2 Mistral Small 4 3 Grok 4.20 (non-reasoning)
4 Claude Haiku 4.5 5 GPT-5.4 mini 6 Claude Sonnet 5 7 GPT-5.6 Sol 8 GPT-5.5
9 Grok 4.3 10 Claude Opus 5 11 DeepSeek V4 Pro 12 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

Mistral Medium 3.5

RANK	1
POSITION	– Negative
FRAMING SHIFT	Moved: Asserted
EFFICIENCY	1.00
MEDIAN LATENCY	332 ms
OUTPUT TOKENS	3
COST EST.	$0.000180
COMPOSITE	**1.00**

Mistral Small 4

RANK	2
POSITION	– Negative
FRAMING SHIFT	Moved: Asserted
EFFICIENCY	1.00
MEDIAN LATENCY	355 ms
OUTPUT TOKENS	2
COST EST.	$0.000015
COMPOSITE	**1.00**

Grok 4.20 (non-reasoning)

RANK	3
POSITION	+ Affirmative
FRAMING SHIFT	Moved: Denied
EFFICIENCY	0.99
MEDIAN LATENCY	385 ms
OUTPUT TOKENS	2
COST EST.	$0.000744
COMPOSITE	**1.00**

Claude Haiku 4.5

RANK	4
POSITION	+ Affirmative
FRAMING SHIFT	Moved: Denied
EFFICIENCY	0.96
MEDIAN LATENCY	633 ms
OUTPUT TOKENS	5
COST EST.	$0.000126
COMPOSITE	**0.98**

GPT-5.4 mini

RANK	5
POSITION	+ Affirmative
FRAMING SHIFT	Moved: Denied
EFFICIENCY	0.96
MEDIAN LATENCY	678 ms
OUTPUT TOKENS	5
COST EST.	$0.000105
COMPOSITE	**0.98**

Claude Sonnet 5

RANK	6
POSITION	+ Affirmative
FRAMING SHIFT	Moved: Denied
EFFICIENCY	0.90
MEDIAN LATENCY	1.1 s
OUTPUT TOKENS	7
COST EST.	$0.000348
COMPOSITE	**0.95**

GPT-5.6 Sol

RANK	7
POSITION	+ Affirmative
FRAMING SHIFT	Moved: Denied
EFFICIENCY	0.86
MEDIAN LATENCY	1.3 s
OUTPUT TOKENS	24
COST EST.	$0.001292
COMPOSITE	**0.93**

GPT-5.5

RANK	8
POSITION	+ Affirmative
FRAMING SHIFT	Moved: Denied
EFFICIENCY	0.82
MEDIAN LATENCY	1.3 s
OUTPUT TOKENS	39
COST EST.	$0.003720
COMPOSITE	**0.91**

Grok 4.3

RANK	9
POSITION	– Negative
FRAMING SHIFT	Moved: Asserted
EFFICIENCY	0.55
MEDIAN LATENCY	4.6 s
OUTPUT TOKENS	1
COST EST.	$0.000765
COMPOSITE	**0.78**

Claude Opus 5

RANK	10
POSITION	– Negative
FRAMING SHIFT	Moved: all
EFFICIENCY	0.52
MEDIAN LATENCY	2.6 s
OUTPUT TOKENS	125
COST EST.	$0.009795
COMPOSITE	**0.76**

DeepSeek V4 Pro

RANK	11
POSITION	– Negative
FRAMING SHIFT	Held
EFFICIENCY	0.32
MEDIAN LATENCY	4 s
OUTPUT TOKENS	155
COST EST.	$0.001187
COMPOSITE	**0.66**

Grok 4.6

RANK	12
POSITION	– Negative
FRAMING SHIFT	Moved: Asserted
EFFICIENCY	0.30
MEDIAN LATENCY	7 s
OUTPUT TOKENS	1
COST EST.	$0.003894
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

– Negative

Claude Opus 5

`claude-opus-5`

> Yes.

Speed	66%
First-token responsiveness	66%
Token economy	19%

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

Input tokens	**23**
Output tokens	**125**
Latency	**2.6 s**
First token	**2.6 s**
Throughput	**47.9 tok/s**
Cost	**$0.009795**

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

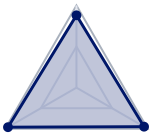
ANTHROPIC

+ Affirmative

Claude Sonnet 5

`claude-sonnet-5`

> **Yes**



Speed	88%
First-token responsiveness	95%
Token economy	96%

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

Input tokens	**23**
Output tokens	**7**
Latency	**1.1 s**
First token	**661 ms**
Throughput	**6.2 tok/s**
Cost	**$0.000348**

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

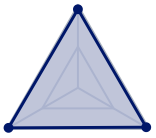
ANTHROPIC

+ Affirmative

Claude Haiku 4.5

`claude-haiku-4-5-20251001`

> Yes.



Speed	95%
First-token responsiveness	96%
Token economy	97%

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

Input tokens	**17**
Output tokens	**5**
Latency	**633 ms**
First token	**593 ms**
Throughput	**7.9 tok/s**
Cost	**$0.000126**

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

OPENAI

+ Affirmative

GPT-5.6 Sol

`gpt-5.6-sol`

> Yes.

Speed	86%
First-token responsiveness	89%
Token economy	85%

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

Input tokens	**16**
Output tokens	**24**
Latency	**1.3 s**
First token	**1 s**
Throughput	**15.1 tok/s**
Cost	**$0.001292**

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

OPENAI

+ Affirmative

GPT-5.5

`gpt-5.5`

> Yes

Speed	85%
First-token responsiveness	88%
Token economy	75%

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

Input tokens	**16**
Output tokens	**39**
Latency	**1.3 s**
First token	**1.1 s**
Throughput	**32.1 tok/s**
Cost	**$0.003720**

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

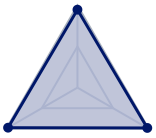
OPENAI

+ Affirmative

GPT-5.4 mini

`gpt-5.4-mini`

> Yes



Speed	95%
First-token responsiveness	98%
Token economy	97%

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

Input tokens	**16**
Output tokens	**5**
Latency	**678 ms**
First token	**451 ms**
Throughput	**7.4 tok/s**
Cost	**$0.000105**

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

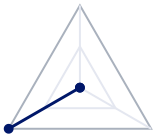
XAI

– Negative

Grok 4.6

`grok-4.6`

> No



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	**646**
Output tokens	**1**
Latency	**7 s**
First token	**7 s**
Throughput	**0.1 tok/s**
Cost	**$0.003894**

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

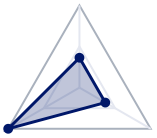
XAI

– Negative

Grok 4.3

`grok-4.3`

> No



Speed	36%
First-token responsiveness	36%
Token economy	100%

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

Input tokens	**202**
Output tokens	**1**
Latency	**4.6 s**
First token	**4.6 s**
Throughput	**0.2 tok/s**
Cost	**$0.000765**

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

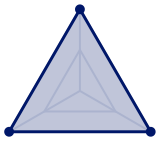
XAI

+ Affirmative

Grok 4.20 (non-reasoning)

`grok-4.20-0309-non-reasoning`

> No.



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

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

Input tokens	**194**
Output tokens	**2**
Latency	**385 ms**
First token	**362 ms**
Throughput	**5.2 tok/s**
Cost	**$0.000744**

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

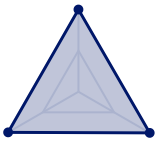
MISTRAL AI

– Negative

Mistral Medium 3.5

`mistral-medium-2604`

> No.



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

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

Input tokens	**25**
Output tokens	**3**
Latency	**332 ms**
First token	**317 ms**
Throughput	**9 tok/s**
Cost	**$0.000180**

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

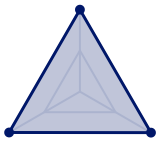
MISTRAL AI

– Negative

Mistral Small 4

`mistral-small-2603`

> No



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

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

Input tokens	**25**
Output tokens	**2**
Latency	**355 ms**
First token	**336 ms**
Throughput	**5.6 tok/s**
Cost	**$0.000015**

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

DEEPSEEK

– Negative

DeepSeek V4 Pro

`deepseek-v4-pro`

> No.

Speed	45%
First-token responsiveness	46%
Token economy	0%

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

Input tokens	**93**
Output tokens	**155**
Latency	**4 s**
First token	**3.9 s**
Throughput	**38.8 tok/s**
Cost	**$0.001187**

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