

20260916-??????-Haiku

thinking OFF thinking ON 11 LLM Claude Haiku

1

- : 2026-09-17
- : Claude (Opus 5)
- : 20260916-ThinkingOFF thinking OFF 20260916-ThinkingON thinking ON
- Haiku : workflow/records/llm_eval_20260916/haiku_answers_20260917.md

0. ???

: Haiku gpt-oss thinking ON 1

BLOCKER 4 3

3

						8	4	4
Claude Haiku	5	1	4	0	1	1	0	3
gpt-oss:120b thinking ON	5	1	4	0	1	1	3	0

	11	1	
Claude Haiku	931 15 31	29.1 169.2	2 49 11
gpt-oss:120b thinking ON	678 11 18	8.4 208.4	11 19 1

Haiku
 [Progress Bar]
 LLM [] 1[Progress Bar]

1. LLM ???

[Progress Bar] : [] Mac [] Anthropic [] API [Progress Bar]
 API [Progress Bar] Claude Code [Progress Bar] Haiku
 [Progress Bar]

1-1. thinking OFF?thinking ON?Haiku ??????

[]	thinking OFF	thinking ON	Claude Haiku[]
[]	[] Mac[]	[] Mac[]	Anthropic []
[]	runner.py [] Ollama[] MLX [] API []	[]	Claude Code [Progress Bar] model: haiku[]
[]	[Progress Bar] []	[]	[Progress Bar] Read []
[]	[]	[]	[Progress Bar] Read [Progress Bar] 2[]
[]	32,768[] 65,536[] 8bit [Progress Bar]	[Progress Bar]	[Progress Bar]
[]	8,192[] 20,000	[Progress Bar] – []	[Progress Bar]
thinking	off	on[] gpt-oss [] medium[]	[Progress Bar]
temperature[] top_p	0[] 1	0[] 1	[Progress Bar]
[Progress Bar]	[Progress Bar] 4 [Progress Bar]	[]	[Progress Bar] 4 [Progress Bar] [Progress Bar]
[Progress Bar]	1[Progress Bar]	[]	[Progress Bar] []
[]	[Progress Bar]	[Progress Bar] []	[Progress Bar] [Progress Bar]
[]	1[]	1[]	11[]
[]	[] 1[]	[] 1[]	[] 1[]

1-2. Haiku ??????

- [Progress Bar] Read
 [Progress Bar]
 [Progress Bar] Web

- 
 - 
 -  4
- 
- 

2. ?????

☐ thinking OFF ☒ thinking ON 11

- `00000000` : `mox_law_v2` `2000` `alexa_handoff_initial` `1000` `mox_law_initial_retry`
`3000` `skm_phase5_compact_retry` `1000` `alexa_weather_initial_retry` `1000`
`smartphone_axis_initial``00000000`
- `0000000000000000` : `mox_law_v3` `airscope_led_initial`
`airscope_led_review2` `codex_control_same_turn_retry`
`every_golf_serigaya_recovery_retry`

3. ?????

 : thinking OFF ☐ thinking ON ☐

3-1. ?????

- $\mathbf{A} : \begin{bmatrix} \text{---} \\ \text{---} \end{bmatrix}$
- $\mathbf{A} : \mathbf{A} \quad 1$
 $\begin{bmatrix} \text{---} \\ \text{---} \end{bmatrix}$
- $\mathbf{A} : \begin{bmatrix} \text{---} \end{bmatrix}$
- $\mathbf{A} : \begin{bmatrix} \text{---} \end{bmatrix}$

3-2. Haiku ?11????????

項目	単位	項目	単位	項目	単位
mox_law_v2	1000 20	項目	49.9%	36,818	項目
alexa_handoff_initial	1000 10	項目	83.1%	44,385	項目
mox_law_initial_retry	1000 30	項目	169.2%	55,616	項目

項目	項目	項目	項目	項目	項目
skm_phase5_compact_retry	項目 1項目	項目	110.1項目	60,380	項目
alexa_weather_initial_retry	項目 1項目	項目	138.1項目	60,373	項目
smartphone_axis_initial	項目	項目	35.5項目	34,229	項目
mox_law_v3	項目	項目	38.2項目	33,698	項目
airscope_led_review2	OK	項目	29.1項目	32,988	項目
every_golf_serigaya_recovery_retry	項目	項目	129.8項目	54,844	項目 PASS項目
airscope_led_initial	項目 BLOCKER項目	項目	37.6項目	34,308	項目
codex_control_same_turn_retry	項目	項目	110.4項目	42,708	項目

3-3. ??????????thinking OFF?thinking ON?Haiku?

項目	項目	gpt-oss OFF	gpt-oss ON	NVFP4 OFF	NVFP4 ON	8bit OFF	8bit ON	Haiku
mox_law_v2	項目 2項目	項目	項目	項目	項目	項目	項目	項目
alexa_han_doff_initial	項目 1項目	項目	項目	項目	項目	項目	項目	項目
mox_law_initial_retry	項目 3項目	項目	項目	項目	項目	項目	項目	項目
skm_phase5_compact_retry	項目 1項目	項目	項目	項目	項目	項目	項目	項目
alexa_weather_initial_retry	項目 1項目	項目	項目	項目	項目	項目	項目	項目
smartphone_axis_initial	項目	項目	項目	項目	項目	項目	項目	項目
mox_law_v3	項目	項目	項目	項目	項目	項目	項目	項目
airscope_led_review2	OK	項目	項目	項目	項目	項目	項目	項目

Item	Category	gpt-oss OFF	gpt-oss ON	NVFP4 OFF	NVFP4 ON	8bit OFF	8bit ON	Haiku
every_golf_serigaya_recovery_retry	Category 1	OFF	ON	OFF	ON	OFF	ON	ON
airscope_initial	Category 2	OFF	ON	OFF	ON	OFF	ON	ON
codex_control_same_turn_retry	Category 3	OFF	ON	OFF	ON	OFF	ON	ON

Category 1 MLX Category 2 thinking ON Category 3 5

3-4. Haiku

- mox_law_v2: Category 1 operation_re Category 2 operation_re Category 3
- codex_control_same_turn_retry: Category 1 same_turn_retry_initial_prompt_mismatch Category 2 106 Category 3
- alexa_handoff_initial: Category 1 PASS Category 2 Hook Category 3 gpt-oss:120b
- mox_law_initial_retry: defect Category 1 PASS Category 2
- skm_phase5_compact_retry: Category 1 PASS FIFO Category 2
- alexa_weather_initial_retry: NO_FINDING Category 1 9 stopped.wav Category 2 SHA 451 Category 3 gpt-oss thinking ON
- every_golf_serigaya_recovery_retry: NO_FINDING Category 1
- 11 BLOCKER Category 1 1 Category 2

4. Claude

Category 1 : Haiku Category 2 8 Category 3

gpt-oss?thinking ON

- Category 1 BLOCKER Category 2 1 Category 3 gpt-oss ON Category 4 4 3

- Mac

gpt-oss?thinking ON????????

- ☐ 

????

- gpt-oss thinking ON

5. ?????????????

-

6. ????????

- haiku answers 20260917.md

7. ????

- duration_ms subagent_tokens [][][][][]

- API `runner.py` Anthropic API `thinking OFF`
- `harness/cases.json` `path`

8. thinking OFF?thinking ON?Haiku ????

`3` `1` Haiku

推理 速度	推理 精度	推理 延迟	推理 吞吐量	推理 成本	推理 精度 推理 延迟	推理 精度	推理 延迟 推理 精度	推理 延迟 推理 精度	推理 延迟 推理 精度
gpt-oss:120b 推理 延迟	推理 精度	4	0	6	0	1	0	4	939推理 延迟 15推理 精度 39推理 精度
gpt-oss:120b 推理 延迟	推理 精度	5	1	4	0	1	推理 精度 1	3	678推理 延迟 11推理 精度 18推理 精度
Qwen3.8 NVFP4推理 延迟	推理 精度	4	1	5	0	1	0	0	547推理 延迟 9推理 精度 7推理 精度
Qwen3.8 NVFP4推理 延迟	推理 精度	4	1	4	推理 精度 1	1	0推理 精度 推理 精度 推理 精度 1推理 精度	1推理 精度 推理 精度	6,079推理 延迟 1推理 精度 41推理 精度 19推理 精度 推理 精度 3600推理 精度 推理 精度
Qwen3.8 8bit推理 延迟	推理 精度	4	2	4	0	1	推理 精度 1推理 精度 推理 精度 推理 精度 1推理 精度	2	1,609推理 延迟 26推理 精度 49推理 精度
Qwen3.8 8bit推理 延迟	推理 精度	1	0	0	推理 精度 5推理 精度 推理 精度 5	0	推理 精度 推理 精度	推理 精度 推理 精度	推理 精度 推理 精度

11 1	11	1	1	1	11 11 11	11	11 11 8 11	11 1 4 11 11	11 11
Claude Haiku 11	11	5	1	4	0	1	11 1	0	931 15 31 11 11 11 2 49

- 11 gpt-oss thinking ON Haiku 5 1
- Haiku NVFP4 thinking OFF 0
- NVFP4 thinking OFF gpt-oss thinking ON
- 8
- Haiku 1

9. 40%30%30%?

40% 30% 30%

LLM LLM

Claude

9-1. ??

		40%	30%	30%	
1	Qwen3.8 NVFP4 thinking OFF	45	87.5	100	74.3
2	gpt-oss:120b thinking ON	55	87.5	80.7	72.5
3	Claude Haiku	55	100	58.8	69.6
4	Qwen3.8 8bit thinking OFF	50	91.3	34.0	57.6
5	gpt-oss:120b thinking OFF	40	75.0	58.3	56.0
6	Qwen3.8 NVFP4 thinking ON	45	80.8	9.0	44.9
7	Qwen3.8 8bit thinking ON	10	61.8	0	22.5

= × 0.4 × 0.3 × 0.3 100

9-2. ????????

- $\text{smartphone_axis_initial} = \text{smartphone_axis_initial} \times 0.5 \div \text{smartphone_axis_initial} \times 10 \times 100$
- $\text{smartphone_axis_initial} \times 0.5 \div \text{smartphone_axis_initial} \times 10 \times 100$
- $\text{smartphone_axis_initial} \times 0.5 \div \text{smartphone_axis_initial} \times 10 \times 100$
- $\text{smartphone_axis_initial} \times 0.5 \div \text{smartphone_axis_initial} \times 10 \times 100$

9-3. ????????

- [redacted]
 - [redacted] 1[] = 1[] :
[redacted]
 - [redacted] 1[] = 1[] : [redacted]
 - [redacted] 1[] = 3[] :
[redacted]
[redacted]
- [redacted] = [redacted] 21[] ÷ [redacted] × 100[]
- LLM [redacted] 0[redacted] Haiku [] 11
[redacted] 0.49[] 2.45[redacted] API [redacted] [redacted] 1[redacted] 5[redacted] 100
[redacted] 490,347[redacted]
- [redacted] : [redacted] Haiku [redacted] API
[redacted] Claude Opus 5 [redacted] 5[redacted] 25[redacted] 100[redacted] 1
[redacted] Haiku [redacted]
Haiku [] gpt-oss[] thinking ON[redacted] 3[redacted] 3[redacted] 1
[redacted] 0.16[] 0.82[redacted] 1
[redacted] Codex
[redacted] 1[redacted]

モデル名	思考モード ON/OFF	思考ステップ数	思考時間 (秒)	推論速度 (tokens/sec)	推論コスト (円/1000 tokens)
Claude Haiku v1.0	0	0	7	21	100
Qwen3.8 8bit thinking OFF	2	0	7	23	91.3
gpt-oss:120b thinking ON	3	0	7	24	87.5
Qwen3.8 NVFP4 thinking OFF	0	0	8	24	87.5
Qwen3.8 NVFP4 thinking ON	1	1	8	26	80.8

9-6. ????????????

- 1 2 1.8
- 3 1 Haiku 69.6
- gpt-oss thinking ON 67.2 2
- 7 8
- Haiku

10. ?????????????????????

: 11

11

10-1. ?????

		thinking			1	
gpt-oss:120b thinking ON	Ollama 11434	think: true = medium	32,768 25,120 131,072	16,384 12,549	600 208.4	131,072 23%
gpt-oss:120b thinking OFF	Ollama 11434	think: false	32,768 65,536 12 15% Claude Code	20,000 off 16,737	600 284.9	12 15% 65,536
Qwen3.8 NVFP4 thinking OFF	Ollama 11434	think: false	32,768 20,264	12,000 8,192 7,070	600 151.3	18GB

配置	模型	thinking 开关	显存 大小	显存 使用	推理速度 (tokens/s)	推理延迟 (ms)
Qwen3.8 NVFP4 thinking ON	Ollama 11434	think: true	65,536 47,387 262,144	40,000 31,850	600 skm_phase5_c ompact_retry 631.4 alexa_weather _initial_retry 3600	262,144 75%
Qwen3.8 8bit thinking OFF	MLX mlx_lm.server 11436	enable_thinking: false	15,550	12,000 8,192 7,014 --max-tokens 1024	600 397.2	28.6GB 56%

11 5 11

10-2. ??

- Ollama
-
-
- OFF NVFP4 46
- 4,096 qwen_shared.py 356 4,096
- 600 Qwen3.8 NVFP4 thinking ON
- 11434 Ollama 11436 MLX 8bit

10-3. ??????????????

- Qwen3.8 8bit thinking ON : 8 50 MLX [metal::malloc] Metal
- Claude Haiku: thinking
- Qwen3.8 bf16:

0000 #4

00 : 2026-09-17 05:34:37 UTC (TRONS-00)

00 : 2026-09-17 06:47:27 UTC (TRONS-00)