NYT-Connections

A Deceptively Simple Text Classification Task That Stumps System-1 Thinkers

Angel Yahir Loredo Lopez[1], **Tyler McDonald[2]**, Ali Emami[2]

[1] - Universidad Autónoma de San Luis Potosí, San Luis Potosí, MX

[2] - Brock University, Niagara Falls, ON, Canada

Connections

FOLD	GAUGE	STANDARD	YARDSTICK
MARY	NORMAL	DRY	BENCHMARK
COMBINATION	CHECK	CALL	MASS
KENT	BET	OILY	WASH

Connections

WASH
DRY
FOLD

	GAUGE	STANDARD	YARDSTICK
MARY	NORMAL		BENCHMARK
COMBINATION	CHECK	CALL	MASS
KENT	BET	OILY	

Connections

WASH

DRY

FOLD

	GAUGE	STANDARD	YARDSTICK
MARY	NORMAL		BENCHMARK
COMBINATION	CHECK	CALL	MASS
KENT	BET	OILY	

Connections

Benchmark	Poker Actions	Skin Types	Starts of U.S States
STANDARD	FOLD	DRY	WASH
YARDSTICK	CHECK	NORMAL	MASS
BENCHMARK	CALL	OILY	MARY
GAUGE	BET	COMBINATION	KENT

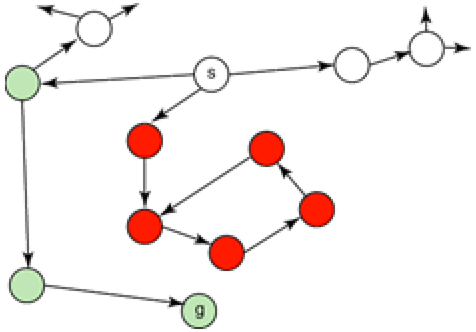
NYT-Connections

- Consists of **554** puzzles designed to **penalize System 1 thinking.**
- Isolated to **text classification** – evaluates **shortcut learning.**
- Updated monthly, providing **novel puzzles.**

Methodology

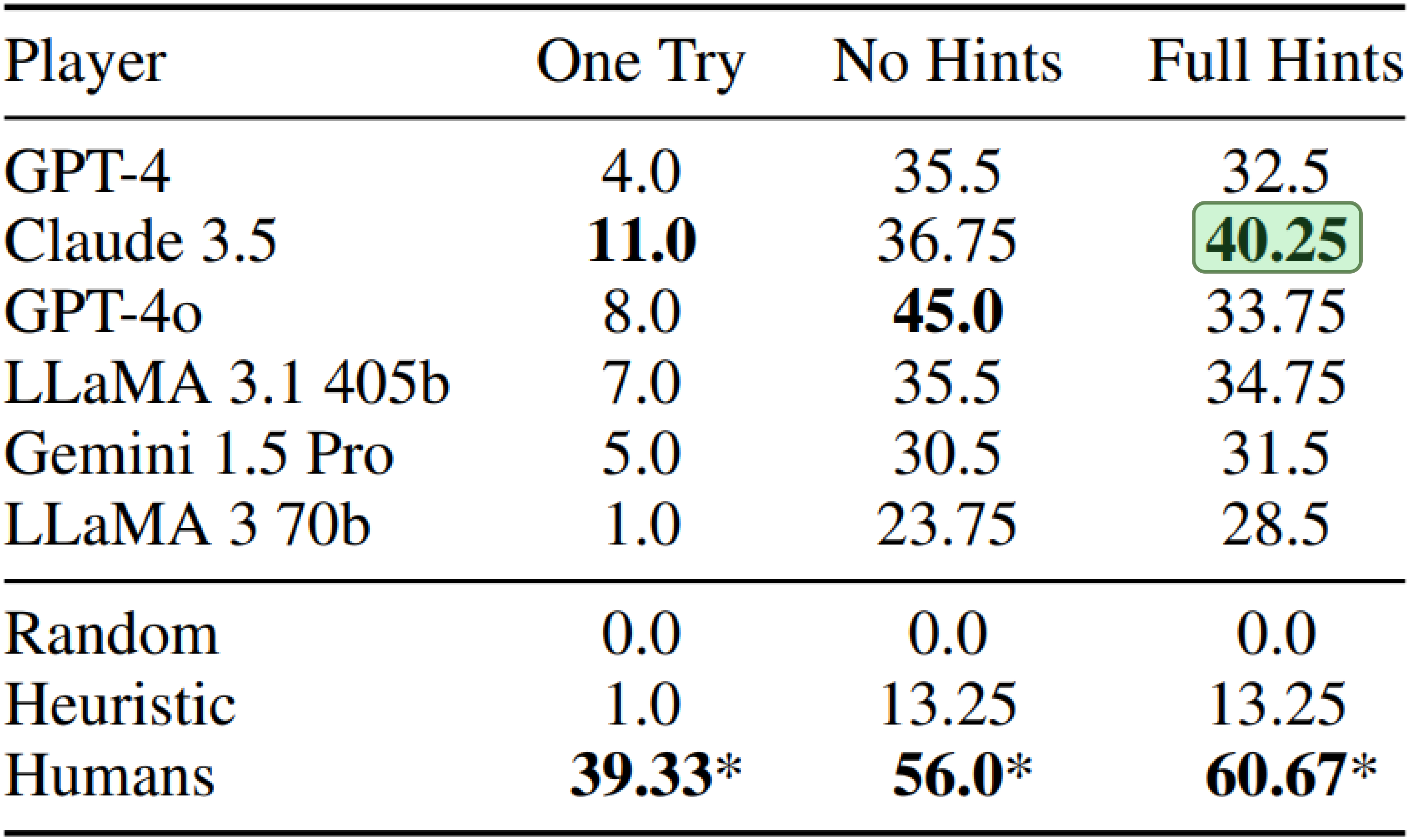
Claude 3.5 Sonnet
GPT-4
GPT-4o
Gemini 1.5 Pro
Llama 3 70B
Llama 3.1 405B

Human Evaluators



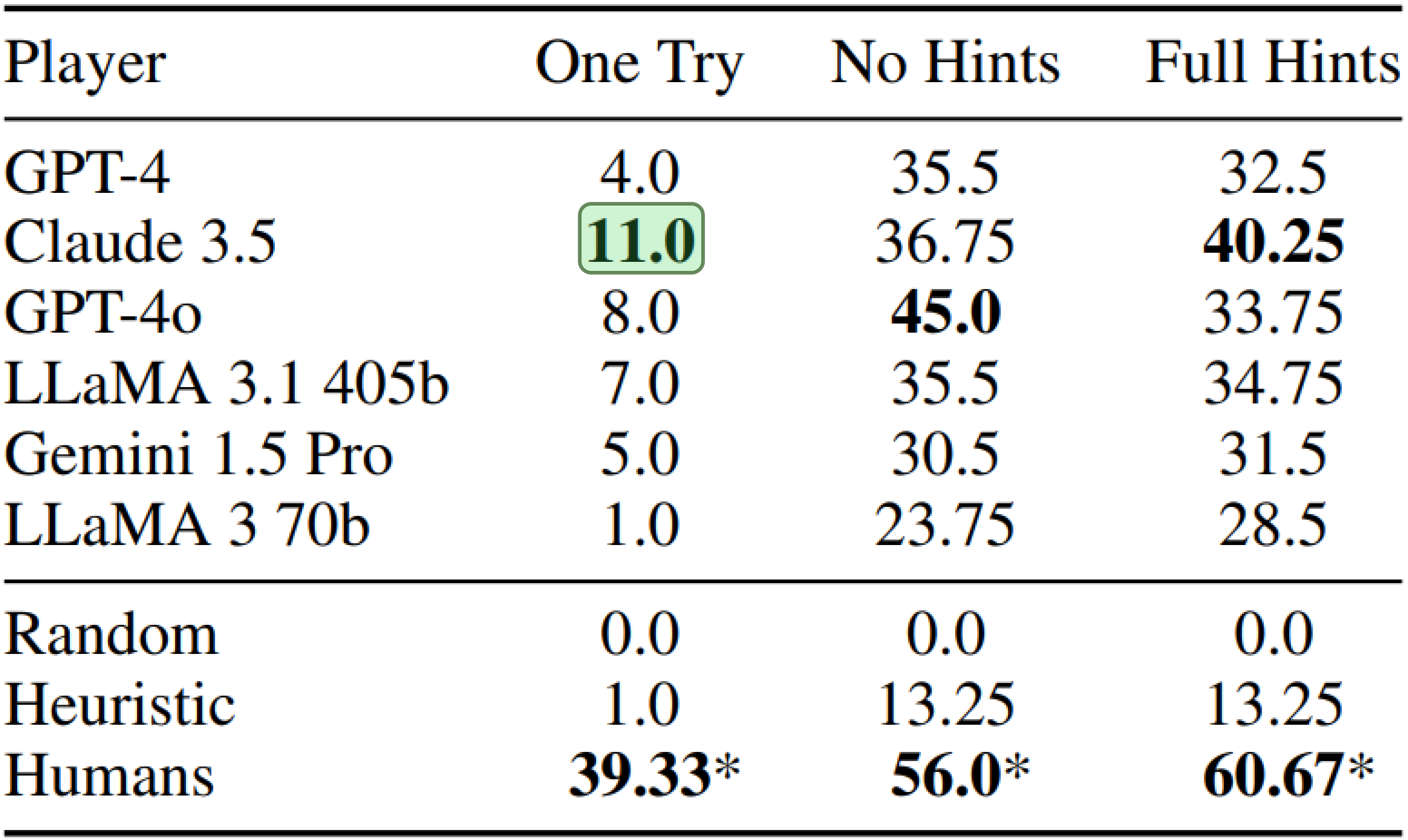
Heuristic Random Guess

Do LLMs Perform Like Humans?



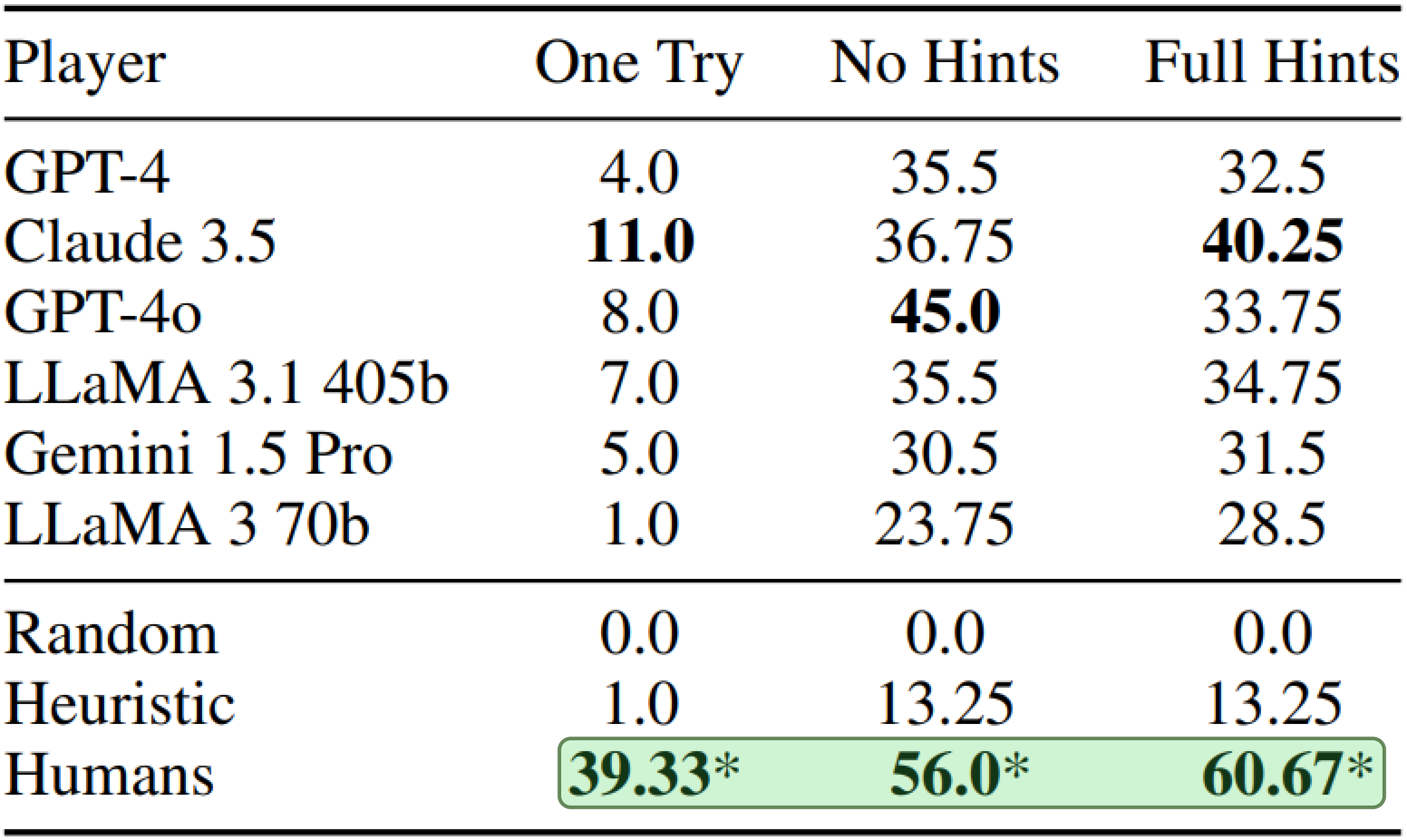
Player	One Try	No Hints	Full Hints
GPT-4	4.0	35.5	32.5
Claude 3.5	**11.0**	36.75	**40.25**
GPT-4o	8.0	**45.0**	33.75
LLaMA 3.1 405b	7.0	35.5	34.75
Gemini 1.5 Pro	5.0	30.5	31.5
LLaMA 3 70b	1.0	23.75	28.5
Random	0.0	0.0	0.0
Heuristic	1.0	13.25	13.25
Humans	**39.33***	**56.0***	**60.67***

Do LLMs Perform Like Humans?



Player	One Try	No Hints	Full Hints
GPT-4	4.0	35.5	32.5
Claude 3.5	**11.0**	36.75	**40.25**
GPT-4o	8.0	**45.0**	33.75
LLaMA 3.1 405b	7.0	35.5	34.75
Gemini 1.5 Pro	5.0	30.5	31.5
LLaMA 3 70b	1.0	23.75	28.5
Random	0.0	0.0	0.0
Heuristic	1.0	13.25	13.25
Humans	**39.33****	**56.0****	**60.67****

Do LLMs Perform Like Humans?



Player	One Try	No Hints	Full Hints
GPT-4	4.0	35.5	32.5
Claude 3.5	**11.0**	36.75	**40.25**
GPT-4o	8.0	**45.0**	33.75
LLaMA 3.1 405b	7.0	35.5	34.75
Gemini 1.5 Pro	5.0	30.5	31.5
LLaMA 3 70b	1.0	23.75	28.5
Random	0.0	0.0	0.0
Heuristic	1.0	13.25	13.25
Humans	**39.33***	**56.0***	**60.67***

Conclusion

- *NYT-Connections* provides a **linguistically isolated and continually novel** challenge requiring **deliberate reasoning**.

- LLMs perform **below human benchmarks**, exhibiting tendencies *between* System 1 and System 2.

- Prompt engineering **is ineffective**, failing to promote System 2 behaviour.

Thank You!

Paper Link

HuggingFace