Language Models Demonstrate the Good-Enough Processing Seen in Humans

Yan Cong and Julia Rayz
(cong4; jtaylor1@purdue.edu)
Purdue University

1 BACKGROUND

Eschers

Escher sentences (comparative illusions):
a language illusion that people accept at first glance but struggle to pinpoint its meaning
(Wellwood et al., 2018; Kelley 2018; Zhang et al., 2023; Zhang, 2024)

1. Weak Escher: *More Brazilians made sandwiches than the American did.*
2. Strong Escher: *More Brazilians made sandwiches than the Americans did.*
3. Canonical comparative: *More Brazilians made sandwiches than Americans did.*

Human sentence processing of Eschers

- *Good-enough*: cognitive efficiency (often at the cost of precision) (Ferreira et al., 2002; Ferreira & Huettig, 2023; Paape, 2024; Kelley, 2018)
- *Rational inference*: prioritizes accuracy through (resource-intensive) error correction (Gibson et al., 2013; Zhang et al., 2023; Paape, 2024)
- Coexistence: flexibility, trade-offs between efficiency and accuracy (Paape, 2024; Brehm et al., 2021; Yadav et al., 2022)

Overall hypothesis

- If LLMs' behavior aligns with humans and can be interpreted through the lens of *good-enough* or *rational inference*,
- LLMs will adeptly alternate between *good-enough* and *rational inference*,
- the dominant processing approach in LLMs will be *rational inference*.

2 Methods

Prompt

Acceptability survey

- Human: Kellye (2018); LLMs: 7-point Likert scale

Hypothesis: acceptability rank of interpretable behavior – control grammatical comparatives > strong Eschers > weak Eschers

Interpretation survey

- Human (n=24, *Prolific*) and LLMs:

More Brazilians made sandwiches than {***Weak**: the American/* ***Strong**: the Americans/* ***Control**: Americans*} did. Who made fewer sandwiches?

A. Americans
B. The Americans
C. The American
D. None of the above

Hypothesis: rational processing dominant – choose A.; good-enough dominant – choose the option that matches surface cues

Probability

- Accumulated tokenwise LLM-surprisal, adjusted for sentence length
- Surprisal of the target word *did* (c.f., Kelley, 2018, EEG signal)

Hypothesis: sensitivity to Eschers of varying illusory strength

Disturbance

- Replaced key nouns with nonsense words (Misra, Rayz, & Ettinger, 2023)
- Replace N before *did* (s1); replace N after *more* (s2); apply both (s1s2)
- Target word *did* prompt + probability

Hypothesis: if LLMs rely solely on syntax – remain consistent regardless of substitution; otherwise, both lexical and syntactic cues matter

3 Results

Prompt

Table 1: Descriptive statistics of prompting GPT-4o-mini in Eschers acceptability ratings, contrasted with human judgments inferred from histograms in Kelley (2018).

GPT-4o-mini	Mean (Std)	Range
Control	5.550 (0.605)	[4, 6]
Filler (Bad)	3.050 (1.146)	[2, 6]
Filler (Good)	5.425 (0.675)	[4, 7]
Strong	5.500 (0.607)	[4, 6]
Weak	5.100 (0.641)	[4, 6]
Human Ratings	**Mean (approx.)**	**Range**
Control	6.5	[1, 7]
Strong	5.5	[1, 7]
Weak	4.5	[1, 7]

Table 2: Comparison between human responses and those from GPT-4o-mini to an Escher-related follow up question, with 112 items for each condition. The number indicates the proportion of choices made.

Condition	Source	NPs	The NPs	The NP
Control	Human	0.761	0.189	0.007
	GPT-4o-mini	0.750	0.250	0.000
Weak	Human	0.008	0.019	0.934
	GPT-4o-mini	0.069	0.181	0.750
Strong	Human	0.040	0.907	0.015
	GPT-4o-mini	0.083	0.889	0.030

Probability

Figure 1: Average tokenwise LLMs-surprisals normalized by sentence length (*surp*).

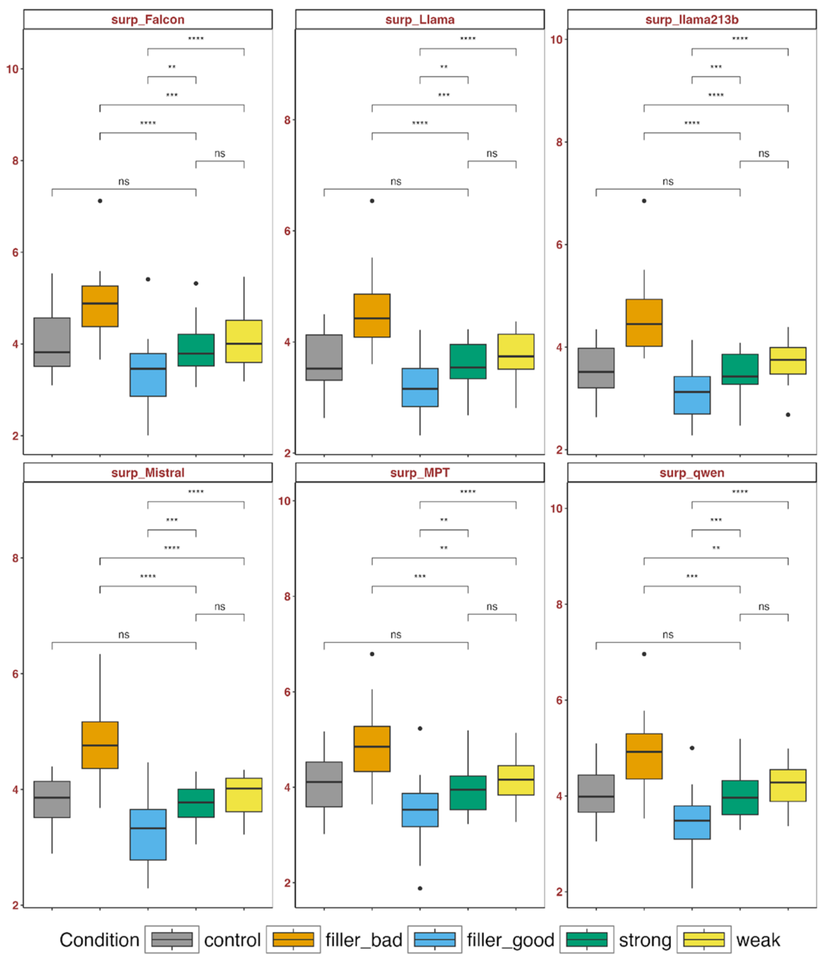
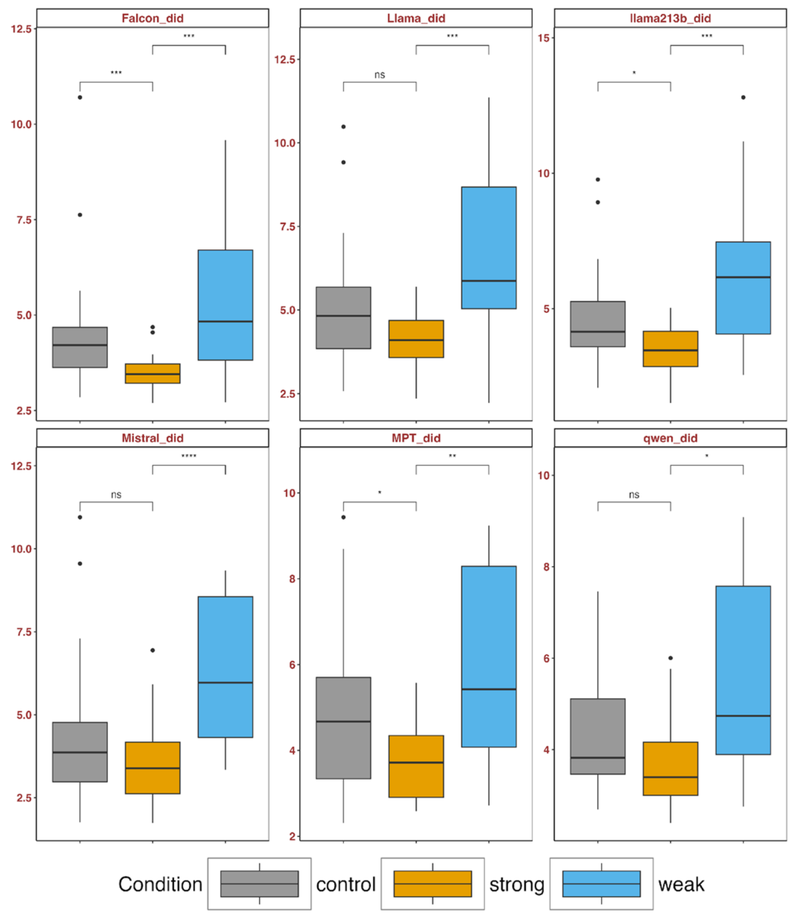


Figure 2: LLMs-surprisals at the target word *did*.



Disturbance

Table 3: Descriptive statistics of prompting GPT-4o-mini in Eschers acceptability ratings on nonce words stimuli.

Condition	Mean (Std)	Range
Control	4.444 (0.836)	[3, 6]
Strong	4.308 (0.766)	[2, 6]
Weak	4.167 (0.768)	[2, 6]

Table 4: Pairwise *Mann-Whitney U* test results for GPT-2 target word *did* surprisal by condition and disturbance type.

Disturbance Type	Comparison of surprisal	*p*-value
s1	Weak > Strong	0.000
s1	Weak > Control	0.006
s1	Strong < Control	0.571
s2	Weak > Strong	0.000
s2	Weak > Control	0.003
s2	Strong < Control	0.004
s1s2	Weak > Strong	0.643
s1s2	Weak > Control	0.732
s1s2	Strong < Control	0.750

4 Discussion

- Prompt: LLMs aligned with humans
- Probability: LLMs struggled to align
- Disturbance: LLMs influenced by both lexical and syntactic cues

LLMs align with the good-enough processing seen in humans.

- Eschers as useful diagnostic tools for probing LLMs' functional linguistic capabilities
- Limitations in current LLMs: robust structure reconstruction and rational error correction

Acknowledgements

We thank Victor Ferreira for the helpful discussion, and Yue Li and Shaohua Fang for sharing the poster template. This research was funded by the College of Liberal Arts, Start-up Funds. Scripts can be found in GitHub https://github.com/yancong222/EschersLLMs.

SCAN ME!