Can language model surprisal predict acceptability and satiation?

Jonathan Merchan Lian Wang Jiayi Lu Judith Degen

{jmerchan, lianwang, jiayi.lu, jdegen}@stanford.edu

Stanford University

Introduction

- Overall, we explore correlations between probabilistic language models and human acceptability judgments
- [1] shows that SLOR predicts human acceptability ratings

$$\text{SLOR} = \frac{\log p_m(\xi) - \log p_u(\xi)}{|\xi|} \tag{1}$$

where $p_m(\xi)$ is the probability assigned to sentence ξ by the LM, $p_u(\xi)$ is its unigram probability, and $|\xi|$ is its length

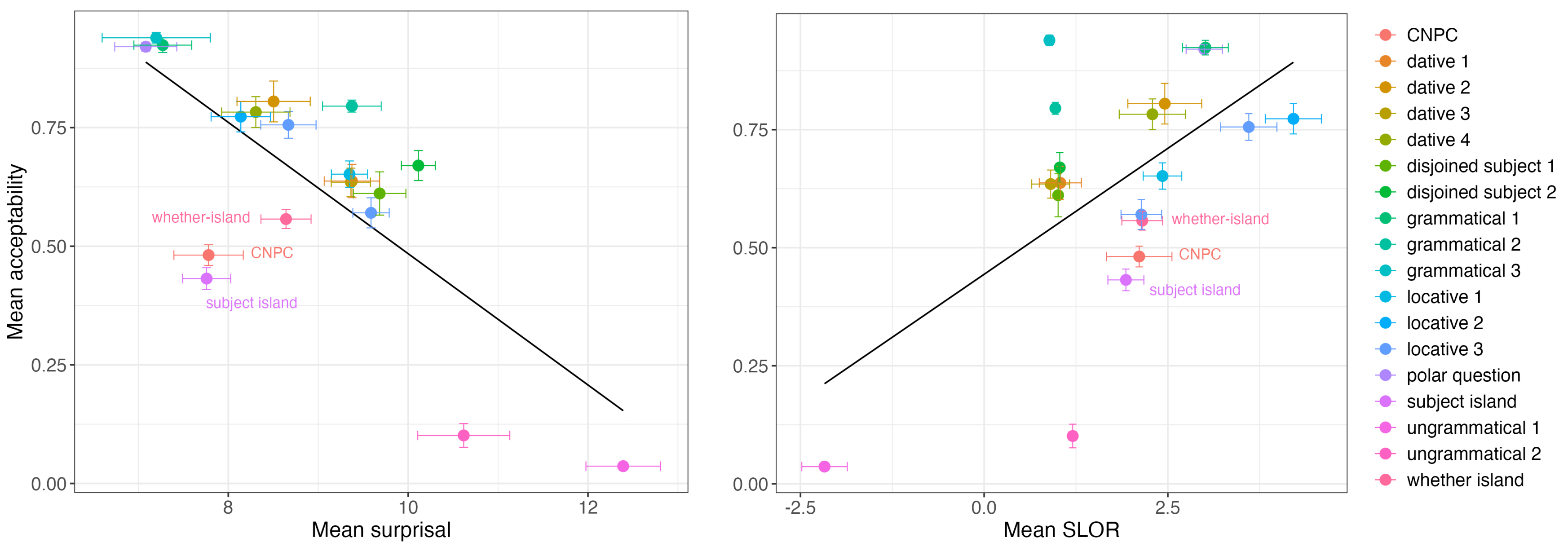
- Exp 1 revisits [1] using a more up-to-date LM (GPT-2 Small) and a controlled variety of sentences
- Exp 2 further explores how change in surprisal after additional training correlates with change in acceptability throughout repeated exposure (the 'satiation effect')

Exp 1: Does surprisal predict acceptability?

- Stimuli and acceptability ratings from [2, 3, 4]; surprisal from pretrained GPT-2 Small

Condition	Example
Polar question	Does the teacher think that the boy found a box of diamonds?
Subject island	What does the janitor think a bottle of ___can remove the stain?
Whether-island	What does the tourist wonder whether the lion attacked ___?
Complex-NP island (CNPC)	Who does the king believe the claim that the prince envied ___?
Dative mismatch*	Kevin gave the children toys and Maria books to the teachers.
Locative mismatch*	Jacob brushed milk on the pastry and Emily the dough with oil.
Disjunction mismatch*	Either Juan or these teachers are making the decisions.

Table 1. Conditions (types of structures). *: Subconditions not shown in table



(a) Mean length-normalized surprisal against acceptability (−Mean LP in [1])

(b) Mean SLOR against acceptability

Figure 1. Results of Exp 1

Exp 2: Does change in surprisal predict rate of satiation?

Two properties of satiation

1. Different structures satiate at different rates (Exp 2a, Exp 2b)
2. Exposure to one structure generalizes to increased acceptability of a different structure (Exp 2c)

Exp 2a: Comparing change in surprisal to rate of satiation

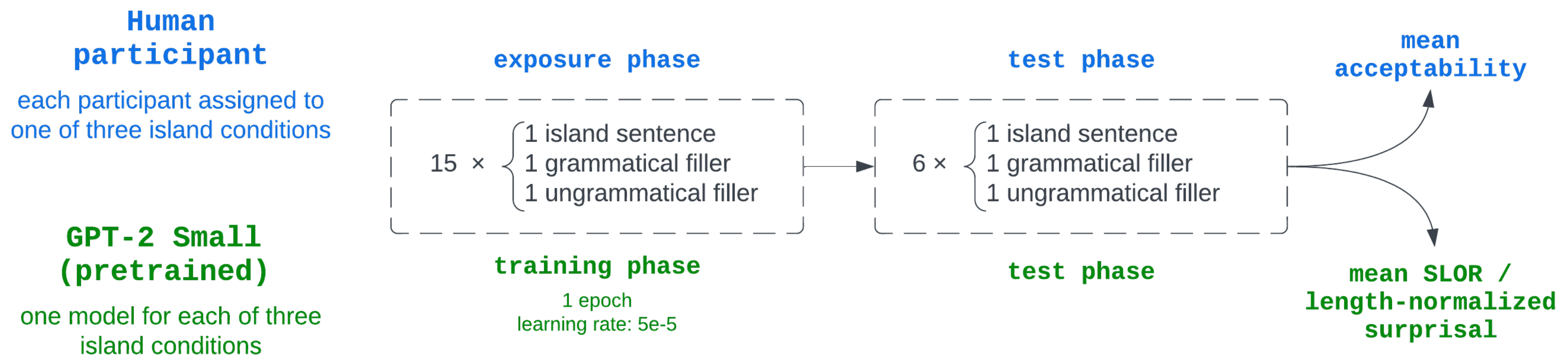
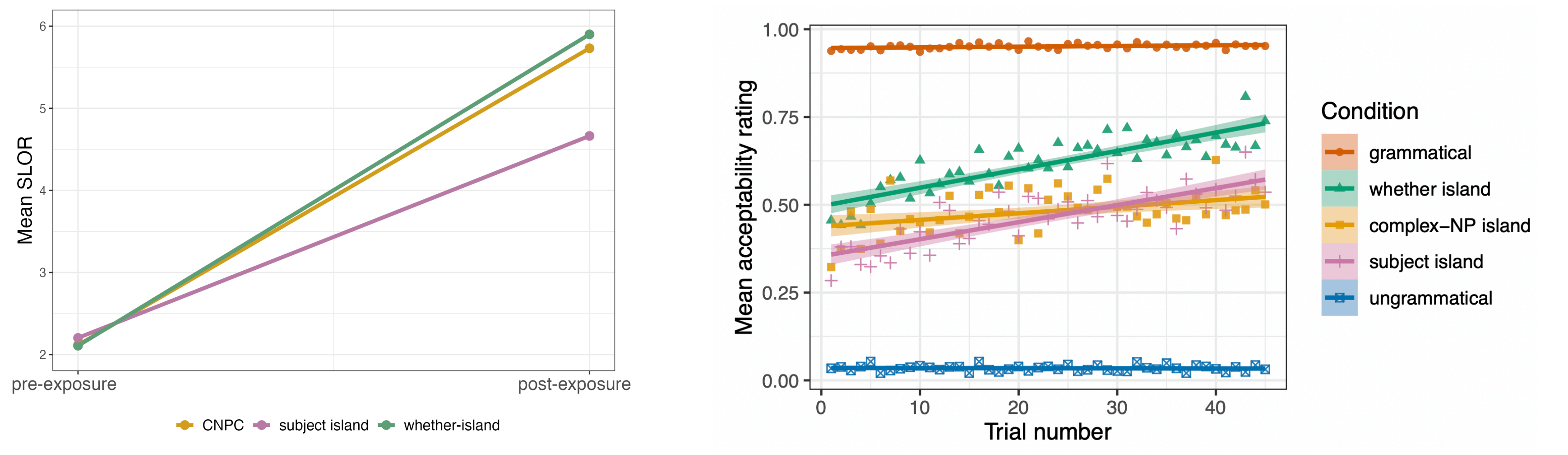


Figure 2. Experimental design of Exp 2a, replication of human experiment in [3] with GPT-2 Small



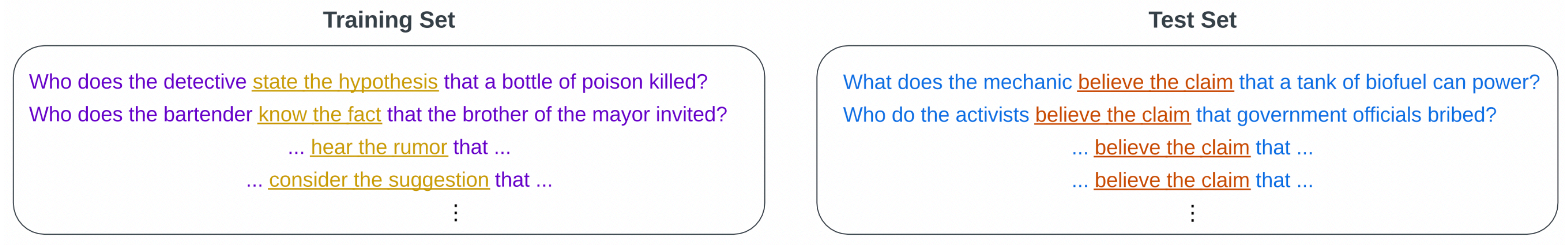
(a) Change in surprisal after fine-tuning

(b) Human satiation results (from [3])

Figure 3. Results of Exp 2a

Exp 2b: Controlling for lexical overlap

- Same design as Exp 2a, but controls for lexical repetition between training and test sets



Training Set

Who does the detective state the hypothesis that a bottle of poison killed?
Who does the bartender know the fact that the brother of the mayor invited?
... hear the rumor that ...
... consider the suggestion that ...
⋮

Test Set

What does the mechanic believe the claim that a tank of biofuel can power?
Who do the activists believe the claim that government officials bribed?
... believe the claim that ...
... believe the claim that ...
⋮

Figure 4. Modified datasets with no lexical repetition between training and test sets

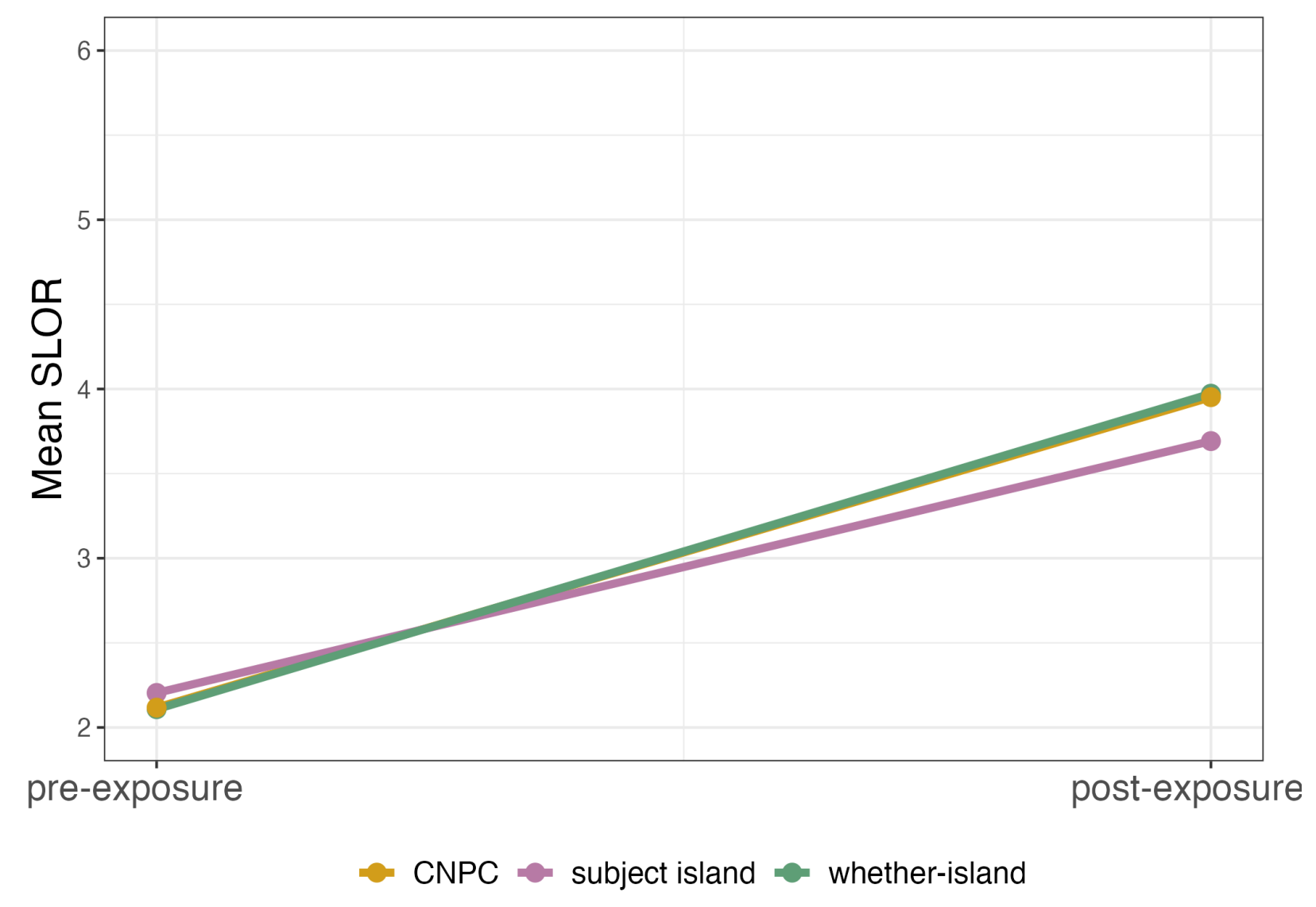


Figure 5. Results of Exp 2b

Exp 2c: Generalization across conditions

- For humans, exposure to subject islands led to increased acceptability of *whether*-islands, but not vice versa (Figs 7a, 7b)
- We test if this holds for GPT-2 Small

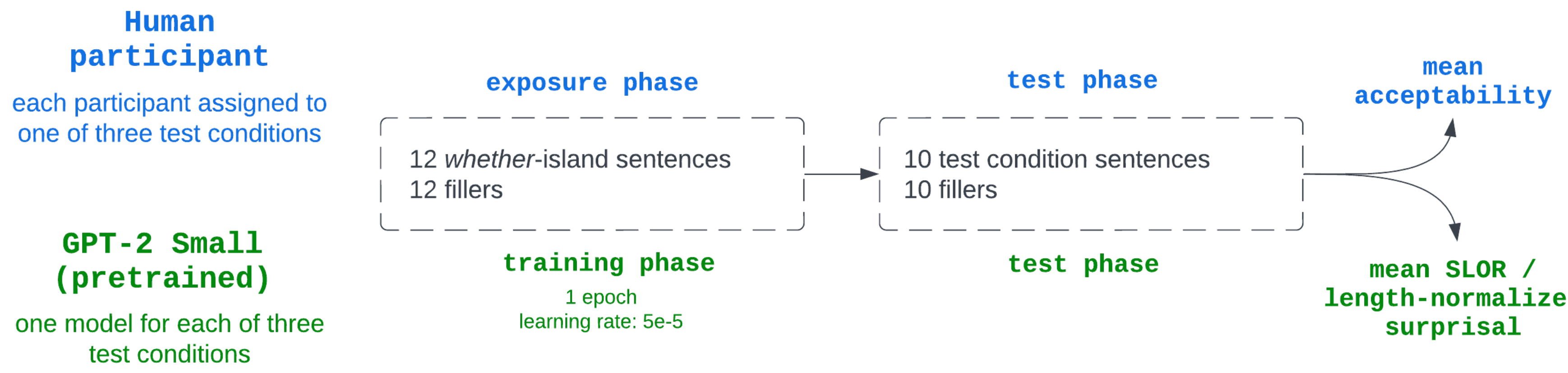
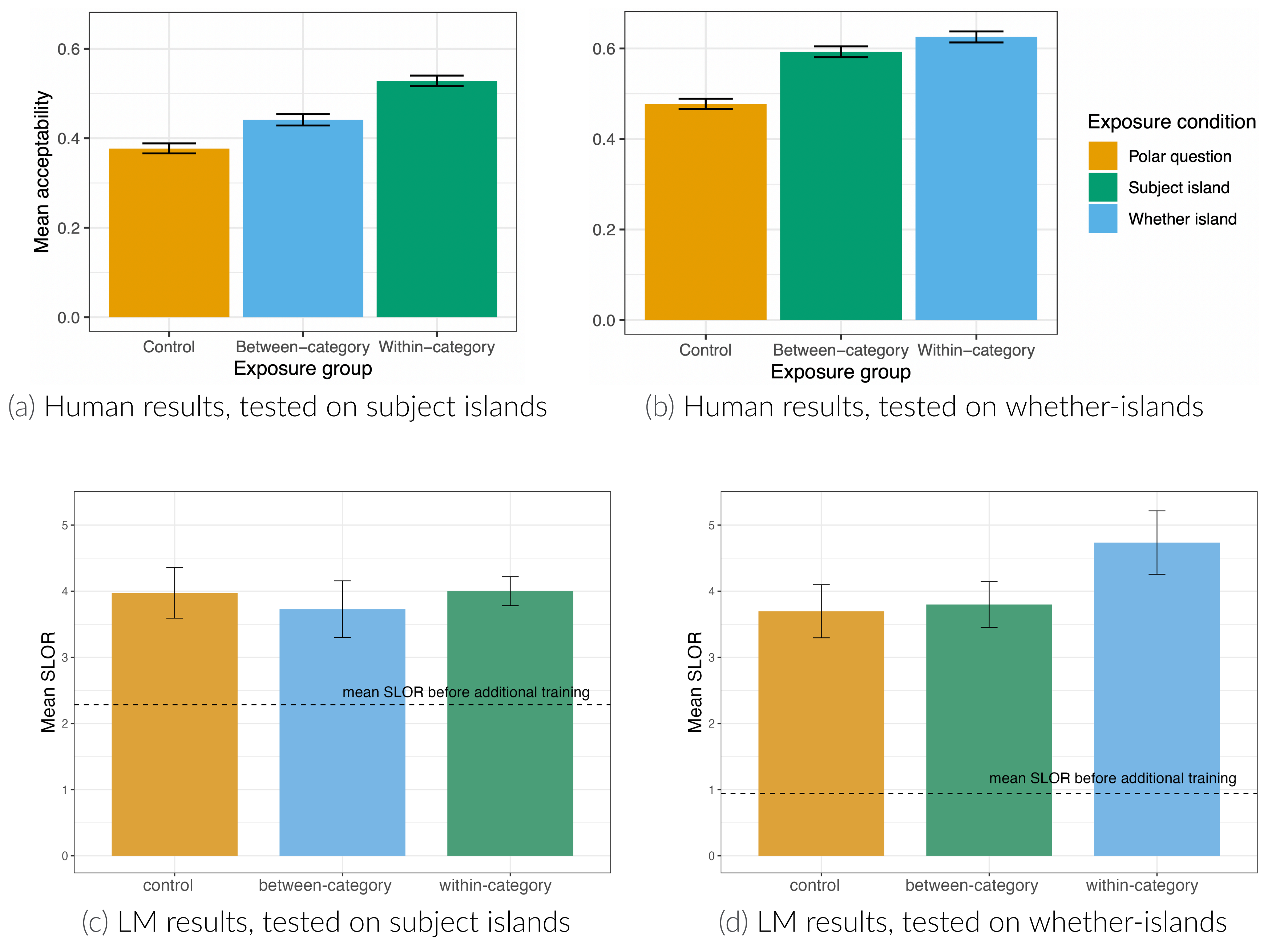

Human participant
each participant assigned to one of three test conditions
GPT-2 Small (pretrained)
one model for each of three test conditions
exposure phase
12 whether-island sentences
12 fillers
training phase
1 epoch
learning rate: 5e-5
test phase
10 test condition sentences
10 fillers
test phase
mean acceptability
mean SLOR / length-normalized surprisal

Figure 6. Experimental design of Exp 2c, replication of human experiment in [4] with GPT-2 Small



(a) Human results, tested on subject islands

(b) Human results, tested on whether-islands

(c) LM results, tested on subject islands

(d) LM results, tested on whether-islands

Figure 7. Results of Exp 2c. Human results from [4].

Discussion

- Surprisal/SLOR generally predicts acceptability
- Fine-tuning models on additional sentences of a structure expectedly leads to decrease in surprisal, analogous to satiation in humans
- However, finer grained discrepancies with human results suggest differences in the relevant representations or learning mechanisms between humans and language models

References: [1] Lau, Clark, and Lappin, *Cognitive Science*, 2017. [2] Lu and Kim, *Frontiers in Psychology*, 2022. [3] Lu, Lassiter, and Degen, *Cognitive Science*, 2021. [4] Lu, Wright, and Degen, *Cognitive Science*, 2022.