Inference and Verbalization Functions During In-Context Learning

Junyi Tao,♦ *Xiaoyin Chen*,♣ *Nelson F. Liu*♦

♦ Stanford University ♣ Mila

TL;DR

Why is LM performance unaffected by **label remappings** during **In-Context Learning**?

e.g., *"true"/"false"* to *"cat"/"dog"*

Using **interchange intervention**, we find:

- 1) LMs use an **inference** function (that solves the task) and a **verbalization** function (that maps the inferred answer to the label space)
- 2) the **inference** function is ***invariant to remappings of the label space***
- 3) these functions are ***located consistently and separately*** in layers

across tasks, datasets, and models (7B-70B)

ICL Setting

DEMO: Sentence 1: {Premise}\nSentence 2: {Hypothesis} \n {Label}

⋮

Sentence 1: {Premise}\nSentence 2: {Hypothesis} \n {Label}

Input-label relation

TEST: Sentence 1: {Premise}\nSentence 2: {Hypothesis} \n

{Label} = true/false —*remap the label space*→ {Label} = cat/dog

Hypotheses

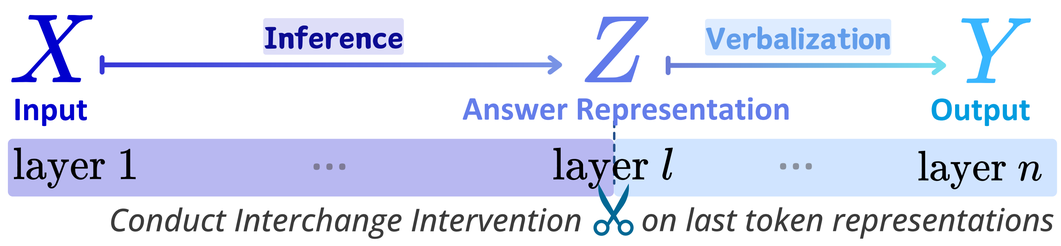
H1	Causal Mechanism	**Inference** Infers the representation of the answer to the task **Verbalization** Maps the answer representation to the label words
H2	Invariance	One shared **Inference** for the same task despite diff. label words
H3	Localization	The functions are separately and consistently localized in the model

Method

*Apply **Interchange Intervention** on representations induced by diff. ICL settings*

Findings

H1 ✓ H2 ✓ H3 ✓



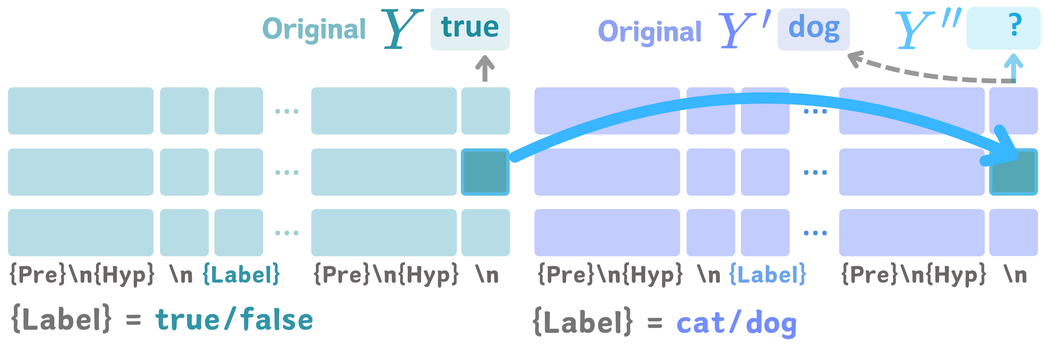
Realize the counterfactual scenarios that would occur only if our hypotheses hold true ✓

Main Experiments: *Remap Label Spaces*

1. Remap label spaces and preserve example-label relations

{Label}=true/false —*remap the label space*→ {Label}=cat/dog

2. Intervene: Interchange representations across settings

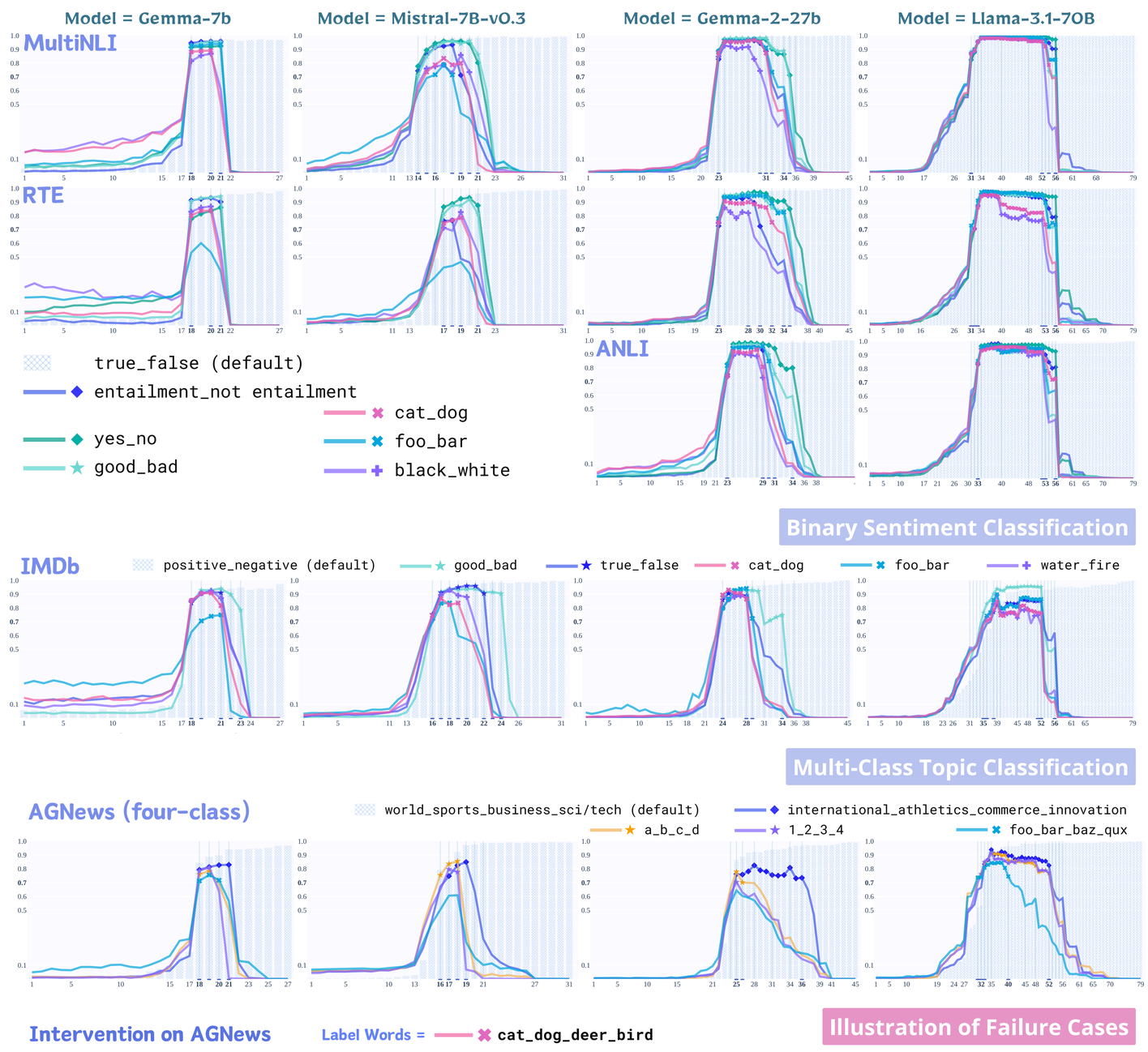


3. Evaluate: Is the output flipped after the intervention?

Y'' flips from dog to Y_c = cat consistently at certain layers

i.e. The intervention flips the answer & preserves the label space

Main Results: Intervention Patterns are Consistent Across Tasks



- These layers **perform** the *causal role* of verbalization → ***H1.2*** ✓
- **Inference** is ***invariant*** *to remappings of* label spaces → ***H2*** ✓
- **Verbalization** is ***consistently*** **located** in late layers → ***H3*** ✓

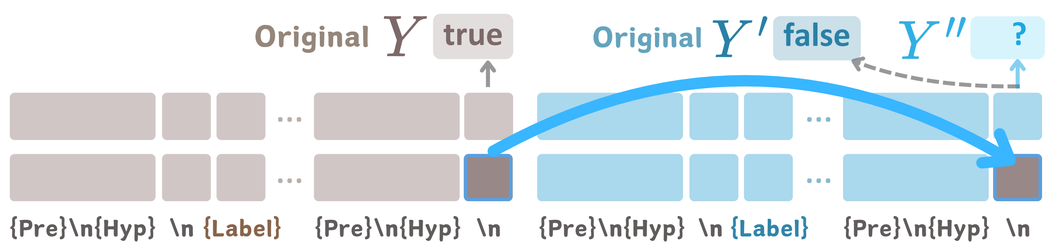
Complementary Experiments: *Reconstruct Tasks*

1. Construct alternative tasks w/ example-label relations

Task = Topic classification, {Label} = true/false

Task = NLI, {Label} = true/false

2. Intervene: Interchange representations across settings

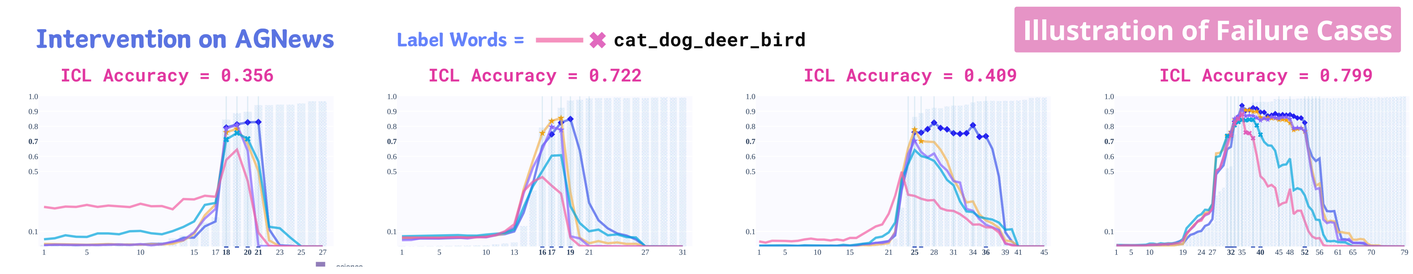


3. Evaluate: Is the output flipped after the intervention?

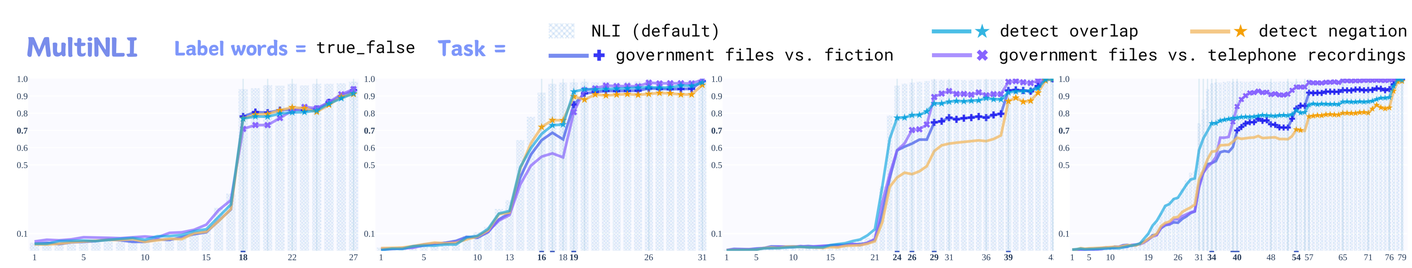
Y'' flips from false to Y_c = true consistently at certain layers

i.e. The intervention flips the answer & preserves the label space

Illustration of Failure Cases



Main Results: Four Reconstructed Tasks on MultiNLI



- These layers **perform** the *causal role* of **inference** → ***H1.2*** ✓