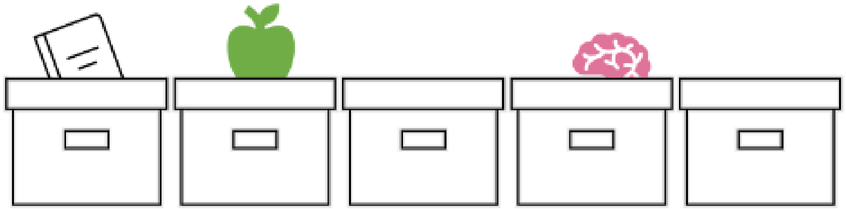
ACL 2023, AREA CHAIR AWARDS, 3 CITATION

Entity Tracking in Language Models

Najoung Kim*
Department of Linguistics
Boston University
najoung@bu.edu

Sebastian Schuster*
Dept. of Language Science and Technology
Saarland University
seschust@lst.uni-saarland.de



What's inside after operations?

Q: Box 1 contains the book. Box 2 contains the apple. Box 4 contains the brain. Move the book into Box 2. Put the bell into Box 4. Move the bell and the brain into Box 5. Box 2 contains _____
A: the apple and the book

[Submitted on 3 May 2023 (v1), last revised 8 Sep 2023 (this version, v2)]

Entity Tracking in Language Models > 简述 > 动机

动机

能够随着语篇的展开，跟踪实体的状态和关系的变化，是长上下文理解和生成连贯文本的先决条件。LLMs 必须能够随着语篇展开准确地表示实体。

Put the eggs, sugar, flour, and baking powder in a bowl and mix to form a light batter. Make sure that the final batter does not contain any lumps of flour or sugar.

油炸挂糊的配置要求

需要

实体识别

- ➢ 识别出 “a bowl” 是引入的新实体；
- ➢ 不应将 “lumps of flour or sugar” 识别成新实体；

共指消解

- ➢ 应当将 “a light batter” 和 “the final batter” 认作相同实体；

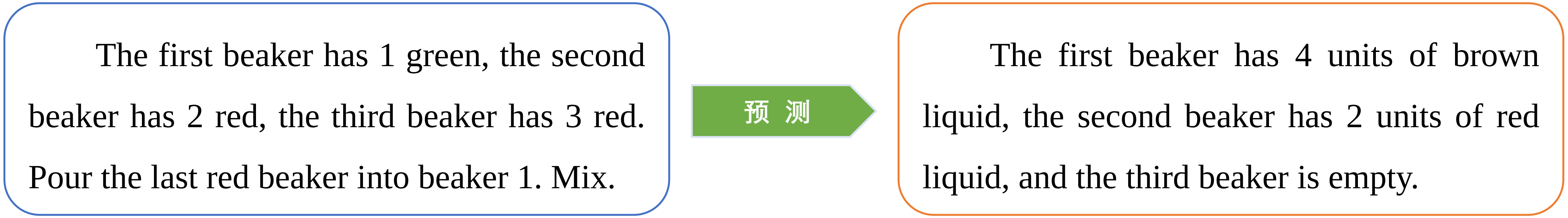
实体跟踪

- ➢ 能够跟踪实体的状态变化；
  如 “鸡蛋已与其他实体混合，不再独立存在”。

Entity Tracking in Language Models > 简述 > 相关研究

相关研究

- 存在测试实体识别、共指消解等问题的数据集;
- 基本没有直接评估 LLMs 跟踪实体状态和关系的的实体跟踪研究。
- 与本论文最接近的工作 —— 2021 年 Implicit Representations of Meaning in Neural Language Models
  - **方法：**对 T5、BART 的模型，根据问题输入的 embedding 进行分类。



The first beaker has 1 green, the second beaker has 2 red, the third beaker has 3 red. Pour the last red beaker into beaker 1. Mix.

The first beaker has 4 units of brown liquid, the second beaker has 2 units of red liquid, and the third beaker is empty.

问题输入

回答输出

- **问题：**
  - 数据集中大部分案例比较 trivial（因 62.7% 的案例始末状态相同、存在 “清空烧杯” 操作）；
  - 其预测任务还需要其他能力（如简单数学运算、颜色混合预测）；

Entity Tracking in Language Models ❯ 简述 ❯ 任务和结论

任务

- **实体跟踪：**给定初始状态的英语描述和一系列状态改变操作，测试 LLMs 推断实体最终状态 的准确性。
- **小模型微调：**在多种文本 (不含代码) 的训练集上微调 T5，研究微调小 LMs 能否推断实体最终状态。

本文结论

- **能行但不完全行：**LLMs 有一定实体跟踪能力。随着操作次数增多，准确率会快速下降。
- **代码预训练有用：**和 GPT-3 对比，经过代码预训练的模型，跟踪实体能力显著提升。
- **微调小模型有用：**在跟踪实体任务上微调 T5，也可以让小模型获得跟踪实体能力。

实体跟踪为主要内容，小模型微调为凑数部分

Entity Tracking in Language Models › 数据集

数据集

基于固定模板的随机采样。

$$W = (O,\ n,\ m,\ e)$$

O	物体，采样自 British National Corpus
n	盒子个数
m	盒子最大容量
e	每个盒子初始状态应当有的物体个数

$$S = \{W,\ Ops,\ NumOps\}$$

W	以 W 为初始状态并更新
Ops	所有可能的操作
NumOps	操作个数，固定值

$O = \{car, apple, bag, toy\}$

$n = 4, m = 4, e \leq 2$

W: Box 0 contains the car, Box 1 contains the apple, Box 2 contains the bag and the toy, Box 3 is empty.

S: Put a coat into Box 3, Move the car from Box 0 to Box 2, Remove the bag in Box 2.

$Ops = \{put, move, remove\}$

$NumOps \leq 3$

Entity Tracking in Language Models > Prompts

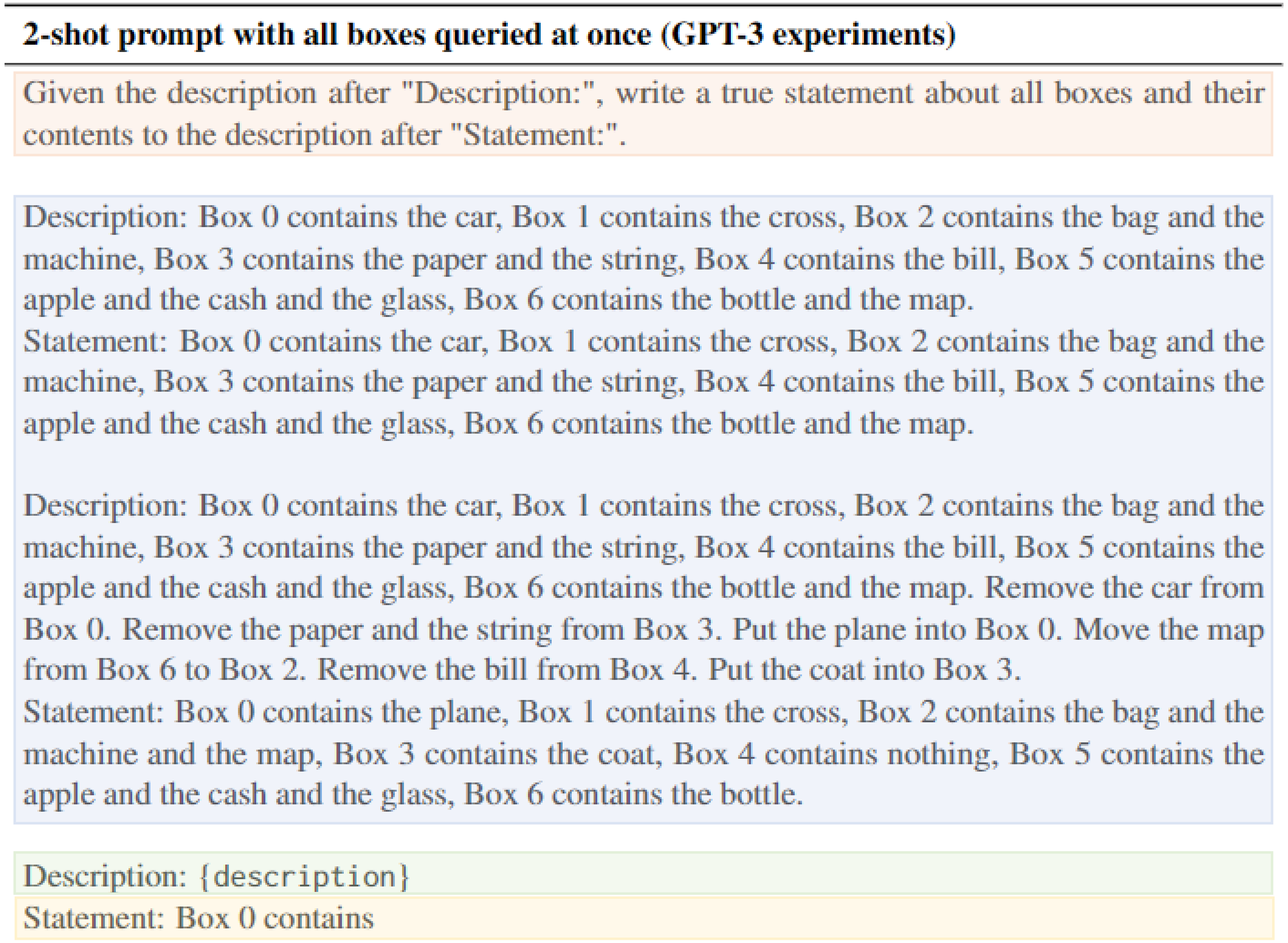
Prompts

General instruction for the task.

Two demonstrations of examples and the expected format.

An initial state description followed by a series of operations.

An incomplete sentence “Box N contains ___”



2-shot prompt with all boxes queried at once (GPT-3 experiments)

Given the description after "Description:", write a true statement about all boxes and their contents to the description after "Statement:".

Description: Box 0 contains the car, Box 1 contains the cross, Box 2 contains the bag and the machine, Box 3 contains the paper and the string, Box 4 contains the bill, Box 5 contains the apple and the cash and the glass, Box 6 contains the bottle and the map.
Statement: Box 0 contains the car, Box 1 contains the cross, Box 2 contains the bag and the machine, Box 3 contains the paper and the string, Box 4 contains the bill, Box 5 contains the apple and the cash and the glass, Box 6 contains the bottle and the map.

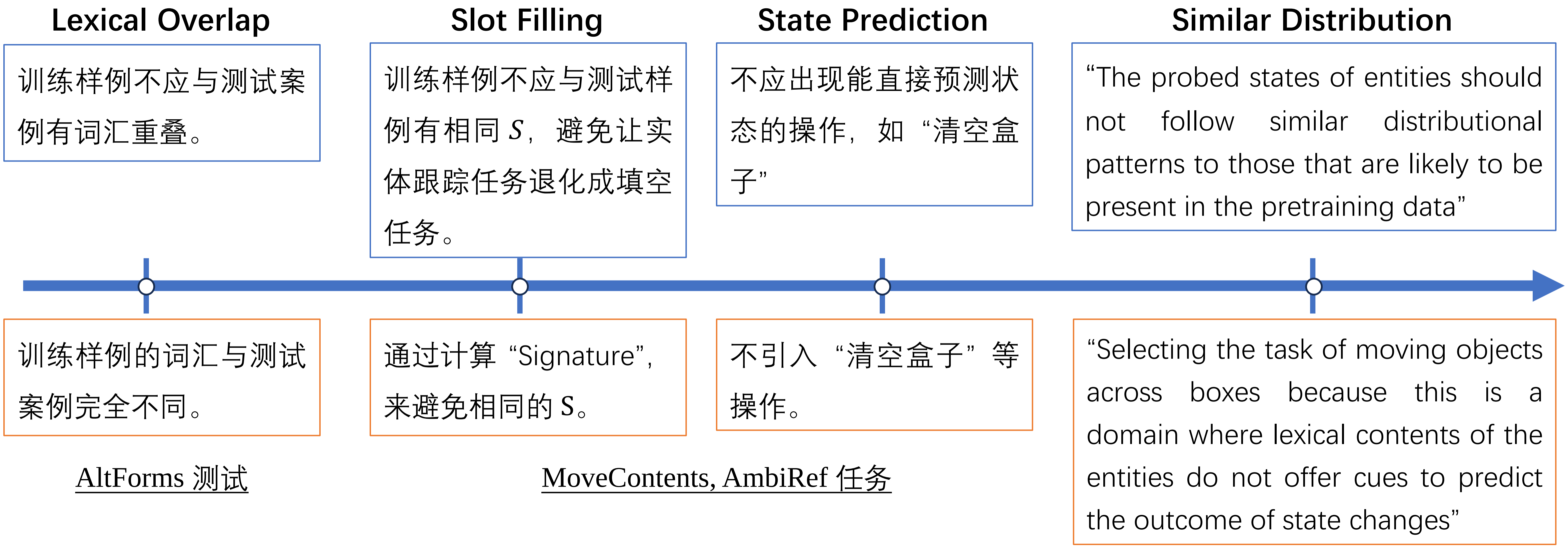
Description: Box 0 contains the car, Box 1 contains the cross, Box 2 contains the bag and the machine, Box 3 contains the paper and the string, Box 4 contains the bill, Box 5 contains the apple and the cash and the glass, Box 6 contains the bottle and the map. Remove the car from Box 0. Remove the paper and the string from Box 3. Put the plane into Box 0. Move the map from Box 6 to Box 2. Remove the bill from Box 4. Put the coat into Box 3.
Statement: Box 0 contains the plane, Box 1 contains the cross, Box 2 contains the bag and the machine and the map, Box 3 contains the coat, Box 4 contains nothing, Box 5 contains the apple and the cash and the glass, Box 6 contains the bottle.

Description: {description}
Statement: Box 0 contains

Entity Tracking in Language Models 数据集的其他考量

数据集的其他考量

测试 LLMs 的跟踪实体能力，不能局限于描述形式、不能根据一些词 / 短语就预测出状态（比如 drain)、不能是常见的状态（比如鸡蛋放在篮子里)。合格的数据集需要避免以下几点。



Lexical Overlap

训练样例不应与测试案例有词汇重叠。

训练样例的词汇与测试案例完全不同。

AltForms 测试

Slot Filling

训练样例不应与测试样例有相同 *S*，避免让实体跟踪任务退化成填空任务。

通过计算 “Signature”，来避免相同的 S。

State Prediction

不应出现能直接预测状态的操作，如 “清空盒子”

不引入 “清空盒子” 等操作。

MoveContents, AmbiRef 任务

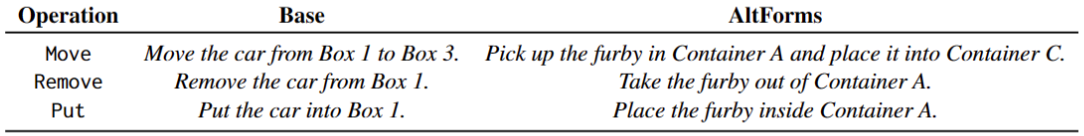
Similar Distribution

“The probed states of entities should not follow similar distributional patterns to those that are likely to be present in the pretraining data”

“Selecting the task of moving objects across boxes because this is a domain where lexical contents of the entities do not offer cues to predict the outcome of state changes”

Experiments - In-context Demonstration

- **AltForms**



Operation	Base	AltForms
Move	*Move the car from Box 1 to Box 3.*	*Pick up the furby in Container A and place it into Container C.*
Remove	*Remove the car from Box 1.*	*Take the furby out of Container A.*
Put	*Put the car into Box 1.*	*Place the furby inside Container A.*

Table 2: Different phrasings of the state-changing operations under the AltForms evaluation setup.

- **MoveContents：**把全部东西从一个盒子移到另一个盒子，避免根据时间顺序和初始描述预测最终状态
- **AmbiRef：**给物体加以修饰词，测试 LLMs 对模糊物体的认知

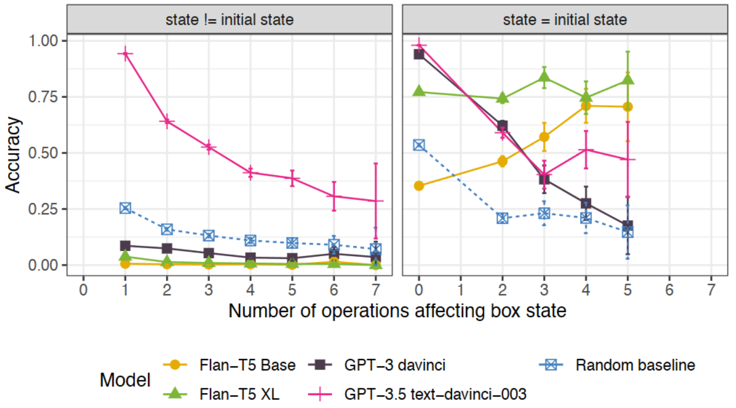


Figure 2: Accuracy on state prediction after n operations that affect a specific box. Left: predictions for boxes whose content differs from the initial state, Right: predictions for boxes whose content is the same as in the initial state. Error bars show 95% CIs.

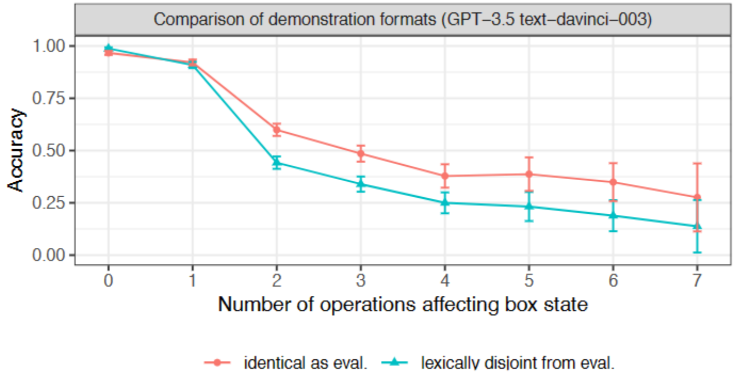


Figure 3: Entity tracking accuracy of `text-davinci-003` with low lexical overlap between demonstration and test examples (AltForms).

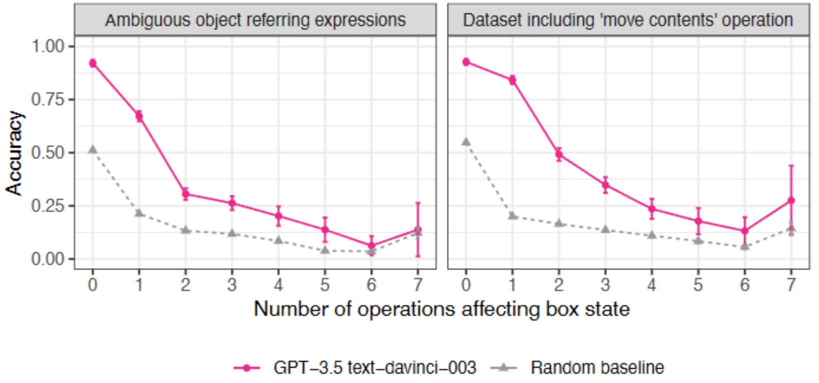


Figure 4: Entity tracking accuracy of `text-davinci-003` for the AmbiRef (left) and MoveContents (right) datasets.

Entity Tracking in Language Models > Experiments - In-context Demonstration

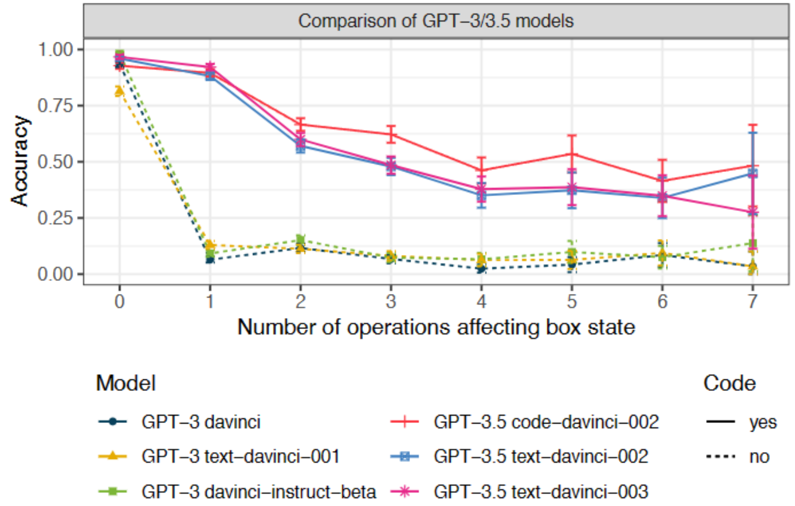
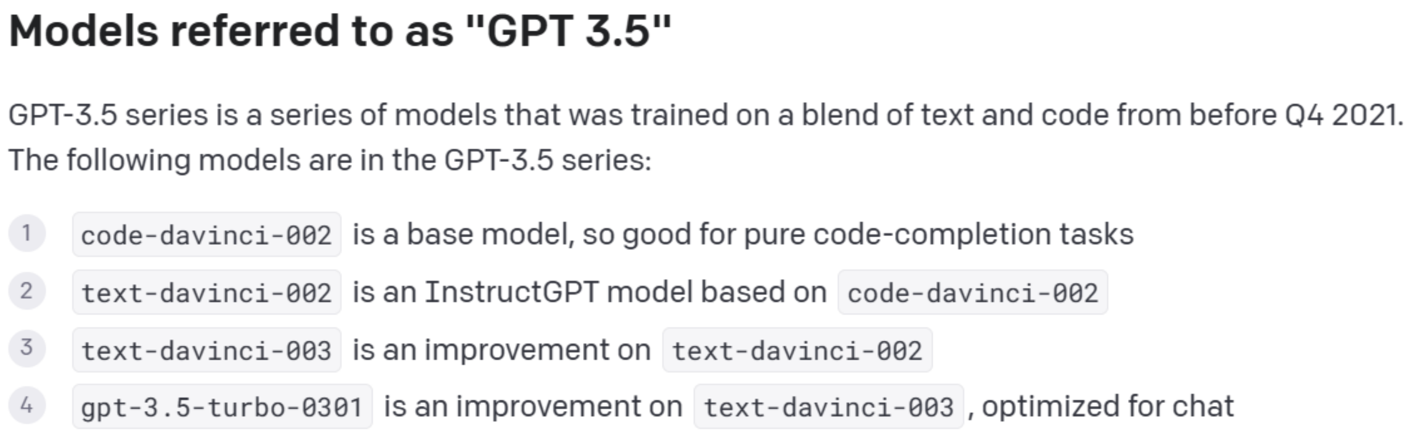


Figure 5: Accuracy on state prediction for different GPT-3 models. Solid lines denote models trained on code and text, and dotted lines denote models mainly trained on text.



Models referred to as "GPT 3.5"

GPT-3.5 series is a series of models that was trained on a blend of text and code from before Q4 2021. The following models are in the GPT-3.5 series:

1. `code-davinci-002` is a base model, so good for pure code-completion tasks
2. `text-davinci-002` is an InstructGPT model based on `code-davinci-002`
3. `text-davinci-003` is an improvement on `text-davinci-002`
4. `gpt-3.5-turbo-0301` is an improvement on `text-davinci-003`, optimized for chat

Entity Tracking in Language Models › Experiments – Fine-tuning

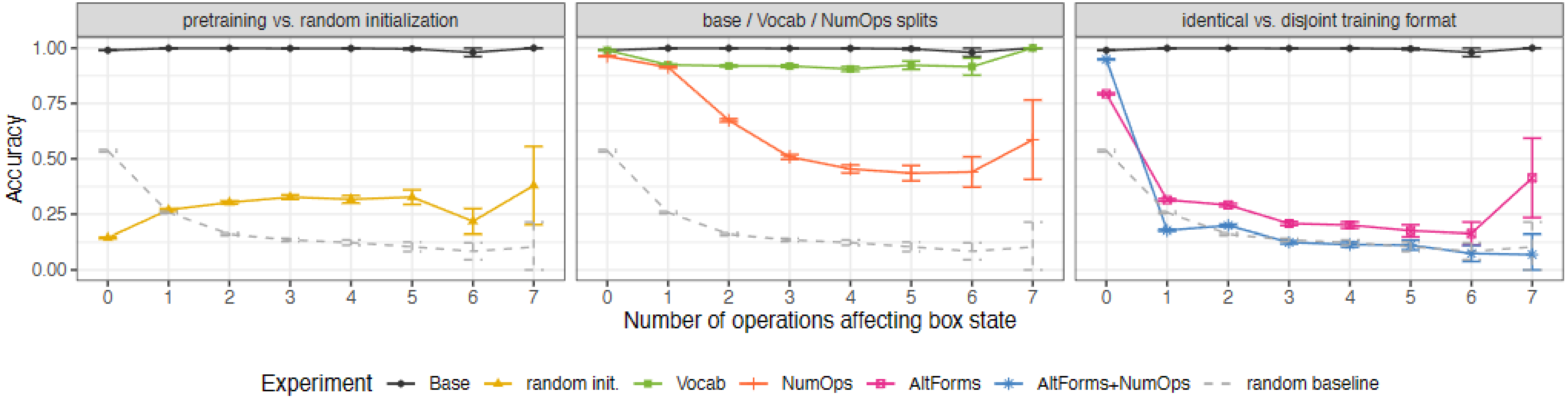


Figure 6: Results for finetuned T5 models.