The Curious Case of Control

Elias Stengel-Eskin
Johns Hopkins University
elias@jhu.edu

Benjamin Van Durme
Johns Hopkins University
vandurme@jhu.edu

Abstract

Children acquiring English make systematic errors on subject control sentences even after they have reached near-adult competence (Chomsky, 1969), possibly due to heuristics based on semantic roles (Maratsos, 1974). Given the advanced fluency of large generative language models, we ask whether model outputs are consistent with these heuristics, and to what degree different models are consistent with each other. We find that models can be categorized by behavior into three separate groups, with broad differences between the groups. The outputs of models in the largest group are consistent with positional heuristics that succeed on subject control but fail on object control. This result is surprising, given that object control is orders of magnitude more frequent in the text data used to train such models. We examine to what degree the models are sensitive to prompting with agent-patient information, finding that raising the salience of agent and patient relations results in significant changes in the outputs of most models. Based on this observation, we leverage an existing dataset of semantic proto-role annotations (White et al., 2020) to explore the connections between control and labeling event participants with properties typically associated with agents and patients.[1]

1 Introduction

Normally-developing children learning English struggle with subject control clauses long after they have successfully acquired the components to understand them (Chomsky, 1969; Cromer, 1970; Maratsos, 1974; Sherman and Lust, 1993). A sentence with a subject control clause has a matrix (main) clause containing a main verb, an agent, and a patient, and an embedded infinitival clause. For example, in *Cole promised Joe to call*, the agent is *Cole*, the patient is *Joe*, and the embedded clause is *to call*. Crucially, the embedded verb here does not have an overt subject, but rather implicitly refers to a subject in the matrix clause (in this case, *Cole*). In a subject control clause, this latent subject of the embedded infinitival clause (usually written as *PRO*) is coindexed with the *subject* (*Cole*) rather than the object (*Joe*) of the matrix (main) clause, i.e. *Cole* (the agent) is doing the calling. This is typically written:

$$[\text{Cole}]_{\text{NP}_i}\ \textbf{promised}\ [\text{Joe}]_{\text{NP}_j}\ \text{PRO}_i\ \text{to call} \quad (1)$$

where subscripts indicate the noun phrase (NP) *"Cole"* is the subject of *"to call"*. (1) can be contrasted with the more common case of object control; for example, if the matrix verb *"promised"* is swapped with an object control verb like *"told"*, then the coreferrent of PRO changes:

$$[\text{Cole}]_{\text{NP}_i}\ \textbf{told}\ [\text{Joe}]_{\text{NP}_j}\ \text{PRO}_j\ \text{to call} \quad (2)$$

Chomsky (1969)[2] finds that children ages 5 to 10 regularly misinterpret subject control (1) for object control (2) while correctly interpreting object control clauses, and proposes that children are following the Minimal Distance Principal (MDP), choosing the linearly closest noun phrase (NP) to govern PRO. Cromer (1970) highlights the systematicity with which children mistake subject control for object control and provides evidence for the MDP. However, Maratsos (1974) argues against the MDP; while his results support the observation that children struggle with subject control, they do not support the MDP, favoring an alternative based on semantic roles. Maratsos changes the subject and object order through passivization:

$$[\text{Joe}]_{\text{NP}_j}\ \textbf{was told}\ \text{by}\ [\text{Cole}]_{\text{NP}_i}\ \text{PRO}_j\ \text{to call} \quad (3)$$

finding that children *correctly* coindex PRO with the (further away) object, violating the MDP.

Recently, large pre-trained language models have shown an impressive ability not only to produce fluent text, but also to perform tasks by "filling

[1] Code and prompts available at https://github.com/esteng/curious-case-of-control.

[2] NB: the author is Carol Chomsky, not Noam Chomsky.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11065 - 11076
December 7-11, 2022

```
You will be given a context and a question. Answer
the question with either "Casey" or "Avery".\n
Context: Avery told Casey to come.\n

Question: Who came, Casey or Avery?\n
Answer:
```

Figure 1: Zero-shot probe for object control. Colors indicate names, which can be swapped.

in the blank" in question-answering prompts, either with no previous examples (zero-shot) or with a few representative examples (few-shot) (Brown et al., 2020; Raffel et al., 2020; Sanh et al., 2021). In light of the difficulty children have in acquiring subject control constructions, we explore how the outputs of the language models tested compare with adult and child strategies for coindexing PRO. In Section 3.2, we examine this question in the zero-shot setting where we give the models only a single question, treating each model as a sort of experimental subject (cf. Fig. 1). Our initial hypothesis is that model outputs will be consistent with child strategies, i.e. the models will perform well on object control examples, but misinterpret subject control for object control. This is informed by two factors: object control is orders of magnitude more frequent than subject control (cf. Section 3.2), and active object control (i.e. (2)) requires resolving a shorter dependency than subject control. We instead find that the tested models fall into three groups, with the majority in fact producing responses mistaking object control for subject control – the opposite of what children do.

Following these observations, in Section 4 we examine to what degree this behavior is sensitive to semantic roles, following Maratsos (1974). To test this, we investigate a "few-shot" setting, where we prompt the model not only with the context and a single question, but also with a set of question-answer pairs that raise the salience of the matrix agent and patient (cf. Fig. 2).

```
You will be given a context and a question. Answer the
question with either "Avery" or "Casey".\n

Context: Avery told Casey to come.\n

Question: Who was told to come, Avery or Casey?\n
Answer: Casey\n
Question: Who told someone to come, Avery or Casey?\n
Answer: Avery\n

Question: Who came, Avery or Casey?\n
Answer:
```

Figure 2: A prompt-hacked example for object control, with long-form instructions.

We find that the models whose behavior in the zero-shot setting was consistent with a positional heuristic have significant differences in the few-shot setting, and the directions of these difference are consistent with a sensitivity to semantic roles.

Finally, in Section 5.2 we investigate whether the sensitivity to semantic roles corresponds to performance on a semantic proto-role labeling task, where models are tested with questions about volition and change of state, properties associated with agents and patients (respectively). We find that while some models are able to perform the labeling task surprisingly well, the differences between models do not necessarily map to the differences in Section 4. We offer three key takeaways:

1. For many models (all GPT-Neo variants, Jurassic Jumbo) the outputs are surprising given their training distributions; while object control is orders of magnitude more common in the text data used in training these models, they perform better on subject control.
2. Large pretrained models are not consistent among themselves. Even models with similar architectures can have very different trends in their outputs, and the outputs of autoregressive and text-to-text models differ substantially.
3. The associations in the autoregressive models tested form outputs that can be explained by simple, often position-based heuristics. For text-to-text models (e.g. T0, T5) the output patterns can also be captured by heuristics based on agent and patient relations. However, sensitivity to agent and patient relations in subject and object control clauses does not always entail higher performance on semantic proto-role labeling.

2 Models and Metrics

We explore both autoregressive models and text-to-text models. Autoregressive models are optimized by minimizing $-\log(P(w_i|w_{-i}))$ for words $w_1, \ldots w_N$ in a given context. These models (also called "decoder-only" models) are composed of just a decoder, which encodes the previously observed tokens $w_1, \ldots w_{i-1}$ produces a probability distribution over the vocabulary for the next token, w_i. Text-to-text models are encoder-decoder models, optimized to reconstruct a noised version of the input via the decoder. An encoder takes a corrupted version of whole sequence $w_1, \ldots w_N$ as input, encoding it into a dense representation from which the decoder reconstructs the original $w_1, \ldots w_N$.

The autoregressive models considered are:

- **GPT-3 Davinci**: this model is only available through the OpenAI API, and its exact training

details are unclear. It is based on the GPT-3 model (Brown et al., 2020) which was trained on Common Crawl (Raffel et al., 2020) with 175B (billion) parameters. Among the several versions of GPT-3, Davinci is generally regarded as the highest-performing (OpenAI).

- **GPT-Neo**: this is an open-source replication of GPT-3 introduced by Black et al. (2021), trained on The Pile (Gao et al., 2020), a 800Gb dataset of web-text intended for pre-training. GPT-Neo has 3 sizes: 1.3B, 2.7B, and 6B parameters (GPT-J), all trained on the same dataset, allowing for direct comparison.
- **Jurassic**: Jurassic Large (7.5B parameters) and Jurassic Jumbo (178B parameters) (Lieber et al., 2021) are also accessible only through an API. The training data is based on Common Crawl, though similarly to GPT-3 Davinci, the details of the training data filtering process are unclear. Relevant differences to GPT-3 are in the tokenization (which includes multi-word expressions) and use of fewer, wider layers.

The text-to-text models we consider are:

- **T5 for QA**: The T5-base text-to-text model (220-million parameters) (Raffel et al., 2020) was pre-trained on cleaned Common Crawl data (C4) and fine-tuned on SQuaD question answering data (Rajpurkar et al., 2016).
- **T0pp**: presented by Sanh et al. (2021), T0pp is an 11B parameter model with a T5-like architecture, pre-trained on Common Crawl data and finetuned specifically for zero-shot question answering on the P3 dataset of NLP benchmarks. This dataset recasts a large number of NLP benchmark datasets into question answering prompts.

Note that because of fine-tuned nature of the "T5 for QA" model, the expected prompt format is fixed, unlike the unrestricted prompt format for the other models. Thus, prompt hacking cannot be done on this version of T5, and so it is only used in Section 3.2. We access non-API models via the Transformers library (Wolf et al., 2020); due to computational constraints, they are run on single GPUs at 1/2 precision. For all models, we decode with the temperature parameter set to 0.

2.1 Metrics

Online APIs make forced decoding very costly (Shin and Van Durme, 2021). Rather than comparing logits for a restricted output vocab, we allow the model to freely generate tokens, letting the model produce a larger variety of answers. In other words, rather than comparing the output probabilities for particular tokens (the logits) given a fixed prefix, we compare full strings of output tokens. However, this method requires heuristics to classify the output strings into categories. In Section 3.2 we validate our heuristics, verifying that for locally-run models the trends are similar when using logits.

Our metric first extracts single word answers and then searches for answers like "The answer is: NAME". For some models and settings, the model re-generates the entire prompt before answering, i.e. it copies the instructions, context, and question, before copying the answer continuation and finally producing an answer. We use Levenshtein distance to check whether the prompt has been regenerated; if it has, it is removed and the first string following the prompt is checked for answer strings. The extraction function always returns the first valid answer produced by the model. If the extraction function fails to find any valid answer strings, the example is skipped in evaluation rather than counted as wrong. We measure significance in model differences with McNemar's test (McNemar, 1947), following Dietterich (1998).

3 Experiment 1

In order to examine what types of generalizations are made by the examined models when prompted for subject and object control information, we construct a number of question-answering-style prompts, where the models fill in the answer. This approach follows recent literature (Raffel et al., 2020; Brown et al., 2020) and takes advantage of the models question-answering abilities. Moreover, using models pretrained with a language-modeling loss rather than training a model specifically for control lets us examine what types of generalizations are captured by the models' associations learned from its original training data, rather than whether a very large model can learn to answer subject and object control questions correctly. We first describe the construction of the prompts used in this set of experiments. Using those prompts, we validate the choice to use heuristic extraction from open generation (i.e. allowing the model to generate tokens up to an end-of-sequence token) rather than logits and confirm that in the C4 dataset, object control is more frequent than subject control. Then, we analyze the zero-shot performance of the models on the subject and object control prompts.

3.1 Subject and Object Prompts

While pretrained language models used for QA are often evaluated in a "few shot" setting, where they are given a few "training" prompts before answering a "test" prompt, in our main experiments we focus on the zero-shot setting. This is in order to avoid learning effects that might result from few-shot prompting (as one would with human subjects) and follows human experiment paradigms, where experimenters are careful not to provide feedback to subjects about the expected answer until after all trials are complete. The prompts used in Section 3.2 and Section 4 have an instruction sentence, a context (like (1)-(3)), a question (e.g. "Who called?"), and an answer continuation. See Fig. 1 for an example zero-shot object control prompt.

We take the max over two instruction types (long and short) in our analyses. Fig. 1 shows the long-form instructions, which include the options in the instructions (e.g. *Answer the question with either "Casey" or "Avery"*) and in the question (e.g. *Casey or Avery*). The short-form instructions omit these prompts in the instructions and questions.

Since the models examined can be sensitive to specific tokens, we cover 9 embedding verbs for object control, chosen from a selection of common linguistics examples: "told", "ordered", "called upon", "urged", "asked", "persuaded", "convinced", "forced", and "pushed". These verbs are presented both in the active (object control experiments) and passive (passive object control experiments). These include verbs that trigger a factuality inference in the affirmative (e.g. "persuaded", "convinced", "forced"). For subject control, we follow previous work (Chomsky, 1969; Maratsos, 1974) and use "promise"; we also include "threaten" as a subject control verb. These verbs are presented only in the active voice, as sentences such as "Casey was promised by Avery to call" were deemed too ambiguous (if grammatical at all). In our main experiments, we use names as NPs; we also report results in Appendix A using common professions to ensure that the trends observed with names hold. We chose 2 male names, 2 female names, and 2 gender-neutral names; these were chosen by taking the top 2 names in each reported gender category in US Social Security data from 1970 to 2019.[3] The gender-neutral names were chosen by taking the top names in the intersection of male and female names.[4] We run the same prompt with each name combination in both orders, to avoid possible biases the model may have towards particular names. When the names are included in the instruction, we add an example with the name order swapped to avoid confounding due the model simply copying the first or last name to appear in the instructions. Finally, for the action infinitive (i.e. the embedded verb) we chose the first 5 coherent verbs (i.e. intransitive infinitives) from a frequency list of English verbs (Yu et al., 2020; Sharov, 2020). This yields 1500 sentences for object control and 150 for subject control (3000 and 300 with swapped names).

3.2 Results and Analysis

Frequency of Subject and Object Control In Section 1, we claimed that object control is more frequent that subject control. To support this claim in the context of the models examined here, we conduct a search of a subset of the C4 dataset (Raffel et al., 2020) for sentences fitting subject control and object control templates. While there are many types of subject and object control, we focus on infinitival complements with transitive matrix verbs, searching with templates similar to the sentences in examples 2 and 1. For object control, we use the same verb list as in Section 3.1. For subject control, we use "promise" and "threaten", as in Section 3.1. We allow the embedded verb to be any verb. We sub-sample the first 1 000 000 sentences of C4 and search them with the templates, finding that object control occurs 10 435 times, while subject control occurs only 209 times, i.e. object control is ~ 50 times more frequent.[5]

Validating Logits Fig. 3 shows the zero-shot accuracy of the best instructions using logit scoring, for the models for which we have access to the full output distribution (non-API models).

Comparing the results to Fig. 4, we see similar but less pronounced trends. As in Fig. 4, GPT-Neo models perform better on subject control and

[3] https://www.ssa.gov/oact/babynames/

[4] We note that this does not guarantee that the name is equally likely for both genders.

[5] Note that object control is to be distinguished from Exceptional Case Marking (ECM; Chomsky et al., 1986) and raising, which yield similar surface forms but have a different analysis than object control. Since object control is already shown to be far more frequent than subject control, we do not examine common ECM and raising verbs; however, we note that, when used with transitive matrix verbs (i.e. verbs with a matrix subject and object) these constructions generally follow the object control template, where the matrix object is the embedded subject.

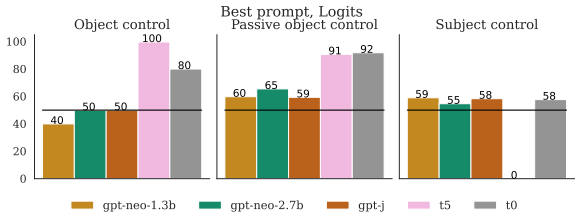

Figure 3: Accuracy of logit-scored model, taking the max across instruction types.

passive object control, while text-to-text models perform better on object control and passive object control. This validates our choice to use heuristics in later experiments. We note also that for GPT-Neo models, the heuristic results in Fig. 4 reflect higher accuracies than the logit-based results in Fig. 3, indicating that the heuristics capture broader range of outputs corresponding to valid answers.

Zero-shot performance In Fig. 4, we see that model classes have different results; we further classify models into 3 groups:

1. The GPT-Neo variants and Jurassic Jumbo are better on subject and passive object control than object control. This pattern can be accounted for by a positional heuristic, namely to take the *first* NP in the matrix clause (i.e. *Maximum* Distance Principle rather than the minimum distance principle of Chomsky (1969)).
2. T5 and T0 are consistent with the observations in Maratsos (1974); both models do better on object control (active and passive) than subject control. This contradicts the MDP but is consistent with a heuristic choosing the matrix patient.
3. GPT-3 and Jurassic Large both perform well above chance on object control and subject control, with their best performance on subject control, but both perform worse on passive object control. This could be matched to a positional heuristic (take the second NP) for object control verbs, rather than an agency-based heuristic.

Further observations Even within model families, there are measurable differences: although GPT-3 and Jurassic Jumbo are roughly the same size and share a general architecture, and are ostensibly trained on similar data, the changes made by Lieber et al. (2021) seem to have a measurable impact, with Jurassic Jumbo performing differently on zero-shot object control examples. For active object control, the difference $\Delta^{\text{active}}_{\text{GPT-3,Jumbo}} = acc_{\text{GPT-3}} - acc_{\text{J-Jumbo}} = 29$ ($p < 0.01$), and for passive $\Delta^{\text{passive}}_{\text{GPT-3,Jumbo}} = -9$ ($p < 0.01$). Similarly, GPT-3 differs from GPT-Neo-1.3B on active object control, and from GPT-Neo-2.7B and GPT-J on both forms of object control, despite sharing an architecture. Further analysis is impeded by a lack of details on the data used to train GPT-3 and Jurassic; this underscores the need for model creators to be transparent about training data and details.

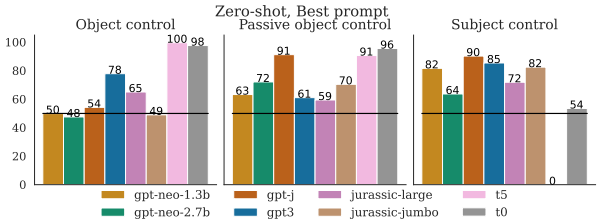

Figure 4: Zero-shot accuracy on object control, passive object control, and subject control. Black line represents random performance (50% accuracy).

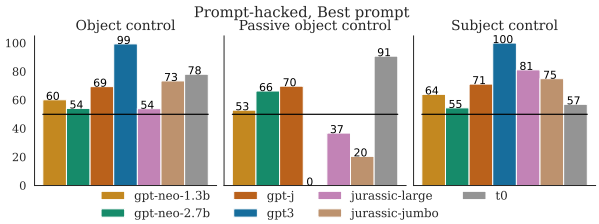

Figure 5: Accuracy on object control, passive object control, and subject control after prompting with agent and patient questions. Accuracy changes from Fig. 4 are generally consistent within heuristic groups.

We also observe that larger models tend to have higher performance: GPT-J is better on all settings than GPT-Neo-1.3B and 2.7B, and Jurassic Jumbo is better than Jurassic Large on passive object control. That said, some larger models are also slightly worse than their smaller counterparts (e.g. Jurassic Jumbo on object control). This suggests that larger models may be more prone to learning patterns corresponding to simple heuristics; however, additional evidence is needed.

4 Experiment 2

Recent work has shown that providing examples in the prompt to a frozen pre-trained model can yield higher performance on QA tasks (Brown et al., 2020). In this setting, typically called "in context learning", some number of demonstrations of questions and answers are given in the context, followed by a test example, to which the model produces an answer. The demonstrations in the context give the model additional guidance on which task is being evaluated. While we do not do in-context learning with training questions about object in subject control, we do experiment with adding information

to the prompt to raise the salience of agents and patients (e.g. "Q: Who told someone to call? A: Cole" for (2).) An example can be seen in Fig. 2; the QA pairs given in the context give the agent and the patient of the matrix clause, but not the embedded clause. This can be thought of as a form of chain-of-thought prompting (Wei et al., 2022) or a scratchpad (Nye et al., 2021) where the gold answers are provided.

We hypothesize augmenting the prompt with agent-patient questions will affect each group as follows:

1. For Group 1 (GPT-Neo, Jurassic Jumbo) where the models' outputs are consistent with positional heuristics, the additional prompts will provide some evidence inconsistent with the heuristic. This evidence may lead to changes in the models' associations that result in outputs less consistent with the heuristics. For example, in Fig. 2 the question *Who was told to come, Avery or Casey?* the answer *Casey* provides evidence against taking the furthest-away NP as the answer. In that case, we would expect a drop in performance in passive object control and in subject control, where the heuristic is beneficial, and an increase in object control, where the heuristic does not help.
2. In Group 2 (T0 and T5), since the model outputs are already consistent with a semantic role-based explanation, we do not expect much change in any setting. In other words, the model outputs are already consistent with access to agent-patient relations combined with an incorrect association for outputting the matrix clause's agent as the embedded subject.
3. Finally, in Group 3 (GPT-3, Jurassic Large), we see that the models' outputs on object control are consistent with an object control-specific heuristic (to take the first NP) the models have lower performance for passive object control than active object control. Thus, as in Group 1, we expect that evidence against the positional heuristic in the prompts will boost performance in passive object control, while reducing it on active object control.

4.1 Results and Analysis

Fig. 5 shows the results after applying prompts with questions about agents and patients. Here, we see that for Group 1 (GPT-Neo and Jurassic Jumbo) the performance does decrease for subject control and passive object control. This decrease is significant for all models and settings ($p < 0.01$) except Jurassic Jumbo in subject control ($\Delta^{\text{subj}}_{\text{Jumbo}} = -7$, $p = 0.06$). At the same time, all object control performance increases significantly for Group 1 ($p < 0.01$). These results confirm our hypotheses, supporting the notion that these models are consistent with a positional rather than semantic heuristic. For Group 2 (T0), we find a significant decrease in performance on both object control types ($\Delta^{\text{active}}_{\text{T0}} = -20$, $p < 0.01$), and no significant difference for subject control ($p = 0.14$); this is roughly consistent with our predictions, since the size of the decrease for object control is relatively small (e.g. compared to the decrease for GPT-3 in the passive). Finally, for Group 3 (GPT-3 and Jurassic Large) we largely see the opposite of what we expected: GPT-3's performance on object control goes close to ceiling after prompting with agent-patient questions, while the passive performance drops to 0; similarly, Jurassic Large's performance drops on passive object control, dropping slightly also on active object control. Both models improve on subject control, with significant differences from the zero-shot setting ($\Delta^{\text{subj}}_{\text{GPT-3}} = 15$, $p < 0.01$, $\Delta^{\text{subj}}_{\text{Large}} = 9$, $p < 0.01$), perhaps reflecting an effect of the semantic role-based priming. Note that for GPT-3, the passive performance drop is from a lack of parseable, non-empty strings being produced, rather than incorrect predictions. For active object control, GPT-3 outputs the second NP more often with additional prompts, increasing the score from 78% to 99%. It may be that the associations responsible for this are also to blame for the degenerate behavior seen in the passive, where the model produces only an end-of-sequence token.

5 Experiment 3

Following Maratsos (1974)'s hypothesis that the observed mistakes children make on subject control sentences is driven by semantic roles, in Section 5.2 we examine the relationship between a model's ability to perform zero-shot object and subject control and its accuracy on identifying attributes commonly associated with agents and patients. Querying language models using fixed semantic role ontologies (e.g. AGENT, PATIENT, THEME) may be difficult as these ontologies may be absent from pretraining corpora. We instead measure the models ability to perform semantic proto-roles labeling (SPRL) for the volition and change of state properties. We use the SPRL data provided in the Uni-

versal Decompositional Semantics (UDS) dataset introduced by White et al. (2020). These properties, first proposed by Dowty (1991), were found to be strongly prototypical of agents and patients, respectively (Reisinger et al., 2015).[6] Proto-role inferences are elicited with simple prompts, circumventing brittle and complicated ontologies. Indeed, the UDS dataset was built by asking annotators questions like *"How likely is it that ARG chose to be involved in the PRED?"* and normalizing their scalar ratings to $[-3, 3]$.

5.1 Semantic Proto-Role Labeling Prompts

To construct a dataset of SPRL prompts, we first filter the UDS dataset for sentences with < 35 tokens – this eliminates many long sentences, which are often more difficult to answer. We then eliminate examples with scalar annotations $\in (-1, 1)$, keeping only examples with strong inferences about the properties. The annotations are binarized with values ≥ 1 leading to a "Yes" answer and values ≤ -1 leading to "No". The annotations are balanced between "Yes" and "No", with the excess examples from the more frequent category being removed.

```
Answer this yes-no question about the following
sentence.\n
Sentence: "Hundreds of people are feared dead in
Mississippi , and the Louisiana city of New
Orleans is badly flooded .”\n
Question: In the event "flooded", does the
participant "city" change in state?\n
Answer: Yes\n
Sentence: "They have unbeatable price in town and
deliver on time .”\n
Question: In the event "have", does the
participant "They" change in state?\n
Answer:
```

Figure 6: Prompt for eliciting SPRL judgments, shown here with one prompting example (1-shot).

Two templates are used for each property; an example template is shown in Fig. 6. For volition, one template asks, '*In the event "PRED", does the participant "ARG" act with volition?*' while the other asks '*In the event "PRED", does the participant "ARG" act on purpose?*'. For change of state, the first template asks, '*In the event "PRED", does the state of the participant "ARG" change?*' and the other asks, '*In the event "PRED", does the participant "ARG" change in state?*' We take the maximum over these two templates.

In our first set of experiments we are interested in the raw ability of the model to perform the semantic proto-role labeling (SPRL) task, and so we allow for full prompt hacking, where demonstrations of the task are provided as part of the context. Accordingly, we stratify the annotations into 4 stages; the bottom stage always forms the "test" prompt, with the answer blank. The remaining 3 stages are added for increasing levels of in-context learning with "training" question-answer pairs (i.e. when the 3rd layer is added, there are 3 example question-answer pairs with answers, and one "test" pair that has no answer, where the model must fill in the answer.) Fig. 6 shows a prompt with one training stage, followed by one test example. We ensure that we use each annotation only once, and that all test annotations paired across across prompting settings. This results in 118 change-of-state test prompts and 168 volition prompts.

Hypotheses We expect that models which perform well on zero-shot subject and object control (e.g. those that can model both the active and passive of object control) will also have higher performance on SPRL, since both require semantic role information. Specifically, we expect to see higher performance from T0 and T5 on at least one of the properties, since their outputs on active and passive object control are consistent with a heuristic that identifies patients as embedded subjects, rather than a positional heuristic. Thus, it may be that the representations learned by T0 and T5 contain more information on agency and patienthood, leading to better performance on SPRL.

5.2 Results and Analysis

Table 1 shows the accuracy on binary semantic proto-role labeling of all models with performance significantly above a random baseline. For change of state, only GPT-3 performs above chance, while for volition, GPT-3, GPT-J, and T0 perform above chance. T0's lower performance is surprising, as the performance of T0 in Fig. 4 is more consistent with an role-based heuristic. In other words, we

Setting	Model	# shots	Acc.	# valid
ΔState	GPT-3	1	0.61	118
Volition	GPT-3	3	0.77	168
	GPT-J	0	0.69	111
	T0	0	0.60	168

Table 1: Accuracy on change-of-state and volition for models significantly above random baseline.

[6] While instigation and stationarity were slightly more predictive of agency and patienthood, they were deemed to be more difficult to re-frame as a prompt.

do not find that models performing well on both passive and active control perform well on SPRL. However, these are separate tasks – thus, it is possible for GPT-3 and GPT-J to be consistent with non-role-based heuristics in one task while still encoding information about agent and patient properties. Finally, we note that in both Fig. 4 and Fig. 5, GPT-3 performs well on subject control and object control in the active, which is consistent with it containing information on agency and patienthood.

6 Related Work

Following the advent of pretrained language models there has been an explosion of work examining what kinds of linguistic knowledge such models contain. Rogers et al. (2020) provide a comprehensive survey of probing work on BERT, covering probing for syntactic, semantic, and world knowledge. This line of probing work generally makes use of linear classifiers on top of frozen representations, tuned on a training set (Adi et al., 2016; Hupkes et al., 2018; Hewitt and Manning, 2019).

In contrast, we follow more recent work in probing large generative models using cloze-style prompts and relatively open generation (Schick and Schütze, 2021). Such models (generative and non-generative) have been probed for diverse knowledge, including syntax (Futrell et al., 2019), symbolic reasoning (Talmor et al., 2020), and commonsense knowledge (Petroni et al., 2019; Kassner and Schütze, 2020; Sakaguchi et al., 2020). This has often been done by recasting benchmark datasets into text, either with zero examples (Sanh et al., 2021) or in the form of in-context learning (Brown et al., 2020; Raffel et al., 2020). Ettinger (2020) present a suite of comparisons between pretrained language models and psycholinguistic experiments. In a similar vein to our work, Lee and Schuster (2022) examine GPT-2's performance on reflexive anaphor agreement in subject and object control clauses, finding that GPT-2 performs well on object control but not transitive subject control; we do not examine reflexive anaphora, and expand our analysis to multiple model classes.

Language models can be sensitive to the format of a prompt – in order to improve extraction of relational knowledge from large language models, Jiang et al. (2020) propose automatic methods for mining new prompts and paraphrasing existing prompts. Similarly, Qin and Eisner (2021) propose a method for gradient-based prompt optimization, and Shin et al. (2020) propose a gradient-based search for prompt token replacement. As our experiments require a specific prompt syntax, we choose to instead run prompts across different instruction styles and name-verb combinations. Large language models are also sensitive to the frequency of terms in their training data; Razeghi et al. (2022) show a strong correlation between frequency of a term in a corpus and performance on tasks requiring that term. This further hightlights how surprising it is that several models perform better on subject control than object control.

7 Conclusion

The results in Fig. 4 indicate that differences between models are not merely of degree, but of kind, with groups of models following different patterns, many of which are inconsistent with the dominance of object control in English. This highlights the pressing need for transparency in the reporting of model details, and especially of training data, without which it is impossible to hypothesize *why* these differences are observed. We also find that, despite there being no trainable parameters in the few-shot setup of Fig. 5, the models tested are in many cases predictably sensitive to semantic role information. Our results in Section 5 suggest that some models appear able to perform semantic proto-role labeling for volition and change of state (Table 1), but this ability is not directly tied to the sensitivity to semantic roles in Fig. 5. In other words, some models contain information on semantic roles but may not recruit that information in producing answers for control examples. This leads to an interesting direction of future work in applying causal mediation analysis (Vig et al., 2020; Elazar et al., 2021) to control clauses, to disentangle the information present in a model from the process by which the model produces an output.

8 Limitations

Firstly, this work is limited by its focus on English syntax, models, and examples. Control constructions exist a variety of languages (Landau, 2001); unfortunately, large pre-trained models currently exist primarily in English alone. Another limitation is the use of fixed prompts: all models tested were found to be sensitive to the prompt format, and while a large number of prompts were explored by varying instructions, names, verbs, and actions, there may be more optimal prompts for the task.

Our work is also limited by the use of open generation. While open-ended generation allows for more flexibility than constrained decoding, it introduces the challenge of interpreting the model outputs, though we do validate the use of open generation in Section 3.2. We note that both these limitations are also common in human subject research.

While our results show that model outputs are consistent with simpler heuristics, some of which are observed in human children, we have attempted to clearly separate this from any anthropomorphic claims that the models might be actively "following" such a heuristic. The claim made is that the associations learned from large-scale pretraining lead to outputs with patterns that can be described by simple heuristics, and that those heuristics at times differ from or resemble heuristics seen in human data. We also note that while we do not make strong commitments to any particular account of human language acquisition, all such accounts differ substantially from how the models tested are trained, i.e. on extremely large text-only corpora.

Acknowledgements

Elias Stengel-Eskin is supported by an NSF Graduate Research Fellowship. This work was supported by NSF #1749025. We would like to thank Madelaine Thomas and Kate Sanders for feedback on previous drafts, as well as the ARR reviewers for their constructive comments and suggestions.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

C. Chomsky. 1969. *The Acquisition of Syntax in Children from 5 to 10*. Research monograph series. MIT Press.

Noam Chomsky et al. 1986. *Barriers*, volume 13. MIT Press (MA).

Richard F Cromer. 1970. "Children are nice to understand": Surface structure clues for the recovery of a deep structure. *British Journal of Psychology*, 61(3):397–408.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800Gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Idan Landau. 2001. *Elements of control: Structure and meaning in infinitival constructions*, volume 51. Springer Science & Business Media.

Soo-Hwan Lee and Sebastian Schuster. 2022. Can language models capture syntactic associations without surface cues? a case study of reflexive anaphor licensing in english control constructions. *Proceedings of the Society for Computation in Linguistics*, 5(1):206–211.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. Technical report, AI21 Labs.

Michael P Maratsos. 1974. How preschool children understand missing complement subjects. *Child Development*, pages 700–706.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

OpenAI. [link].

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning. *arXiv preprint arXiv:2202.07206*.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

S Sharov. 2020. Know thy corpus! robust methods for digital curation of web corpora. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. Leeds.

Janet Cohen Sherman and Barbara Lust. 1993. Children are in control. *Cognition*, 46(1):1–51.

Richard Shin and Benjamin Van Durme. 2021. Few-shot semantic parsing with language models trained on code. *arXiv preprint arXiv:2112.08696*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33:12388–12401.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Aaron Steven White, Elias Stengel-Eskin, Siddharth Vashishtha, Venkata Subrahmanyan Govindarajan, Dee Ann Reisinger, Tim Vieira, Keisuke Sakaguchi, Sheng Zhang, Francis Ferraro, Rachel Rudinger, et al. 2020. The universal decompositional semantics dataset and decomp toolkit. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5698–5707.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Charles Yu, Ryan Sie, Nicolas Tedeschi, and Leon Bergen. 2020. Word frequency does not predict grammatical knowledge in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4040–4054, Online. Association for Computational Linguistics.

A Additional Controls

To ensure that our results are not biased by the use of names, we replicated a round of experiments using common professions rather than names. The professions used were “doctor", “lawyer", “engineer", “writer", “janitor", “bartender", e.g. *“The bartender told the engineer to come.”*

Profession prompts were run only with the long instruction format. In Fig. 7 we include for reference the results from the name experiments shown in Fig. 4 but restricted to only the long prompt format (Fig. 4 took the best accuracy over short and long instructions). Fig. 8 shows that similar trends can be seen when using professions rather than names, confirming that the results are not due to name-specific processing.

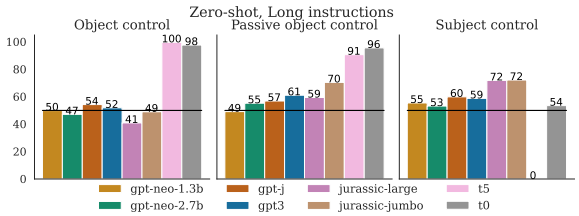

Figure 7: Accuracy of long instruction template on names, for comparison to Fig. 8

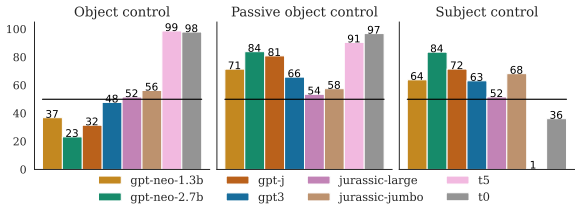

Figure 8: Accuracy of long instruction template on professions. Performance follows similar trends to comparable results with names (Fig. 7).

B Licensing

All data and code is released under an MIT license.