# The Curious Case of Control

**Anonymous ACL submission**

## Abstract

Children acquiring English make systematic errors on subject control sentences (Chomsky, 1969) possibly due to heuristics based on semantic roles (Maratsos, 1974). Given the advanced fluency of large generative language models, we ask what kinds of generalizations these models make on object and subject control clauses. We find broad differences between models, with many models adopting 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 text data.

## 1 Introduction

Normally-developing children learning English have been found to 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). In a subject control clause, the latent subject of an embedded infinitival clause (usually written as *PRO*) is coindexed with the *subject* rather than the object of the matrix (main) clause. For example, in the sentence, *"Cole promised Joe to call"* the complement *"to call"* has a subject not expressed overtly (Cole). 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 *"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) finds that children ages 5 to 10 regularly misinterpreted subject control (1) for object control (2) while correctly interpreting object control clauses. Chomsky 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:

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

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

Recently, large pre-trained language models (PLMs) have shown an impressive ability not only to produce fluent text, but also to perform tasks in zero- and few-shot settings via prompting, especially for question answering (QA) (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 whether the outputs of PLMs are consistent with adult or child strategies for coindexing PRO. We examine this question in the zero-shot setting, treating each model as a sort of experimental subject. 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. Appendix D.1), and active object control (i.e. (2)) requires resolving a shorter dependency than subject control. We instead find that the PLMs tested fall into three behavioural groups, with the majority in fact producing responses that mistake subject control for object control – the opposite of what children do. We show that this behaviour is sensitive to semantic roles, mirroring Maratsos (1974)'s findings. We will release our code and prompts at `http://anonymous_url.com`.

## 2 Methods

**Subject and Object Control** While PLMs used for QA are often given a few "training" prompts (few-shot setting) before answering a "test" prompt, in our main experiments, to avoid learning effects that might result from few-shot prompting (as one would with human subjects), we focus on the zero-shot setting. The prompts used in Section 3 are made of an instruction sentence, a context (like (1)-(3)), a question (e.g. "Who called?"), and an answer continuation. Examples of all prompts are given in D.2. We take the max over two instruction types (long and short) in our analyses. While we do not do "prompt hacking" with training questions about object in subject control, we do experiment with adding QA information to the prompt to raise the salience of agents and patients (e.g. "Q: Who told someone to call? A: Cole" for (2).)

Since the models examined can be sensitive to specific tokens, we cover 9 embedding verbs for object control: "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). For subject control, we follow previous work (Chomsky, 1969; Maratsos, 1974) and focus on "promise" In our main experiments, we use names as NPs; we also report results in Appendix B 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.[1] 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. 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 process yields 1500 sentences for object control and 150 for subject control (3000 and 300 when names are swapped).

**Semantic Proto-Role Experiments** Following Maratsos (1974)'s hypothesis that the observed mistakes children make on subject control sentences is driven by semantic roles, in Section 3 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. Since querying language models using fixed semantic role ontologies may be difficult 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 Universal 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).[2] By design, SPR inferences are elicited with simple prompts, circumventing brittle and complicated ontologies. Indeed, the UDS dataset was created by asking annotators questions like "How likely is it that ARG chose to be involved in the PRED?" with crowdworkers giving scalar ratings normalized to $[-3, 3]$.

To construct a dataset of SPRL prompts, we first filter the UDS dataset for sentences with $< 35$ tokens. We then eliminate examples with scalar annotations $\in (-1, 1)$, keeping only examples with strong inferences about the properties. The annotations are binarized ($> 1$ = "Yes") and balanced across "Yes" and "No". Two QA templates are used for each property (cf. Appendix D.3).

In Section 3 we are interested in the raw ability of the model to perform the SPRL labeling task, and so we allow for prompt hacking. 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 prompting with "training" QA pairs. We ensure that none of the annotations is used more than once, and that all test annotations are the across prompting settings, allowing them to be paired, resulting in 118 change-of-state test prompts and 168 volition prompts.

**Models** We explore both autoregressive models and text-to-text (T2T) models. Autoregressive models are optimized by minimizing $-\log(P(w_i|w_{-i}))$ for words $w_1, \ldots w_N$ in a particular context. These models have just an encoder. T2T models are encoder-decoder models, optimized to reconstruct a noised version of the input via the decoder.

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

[2] 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.

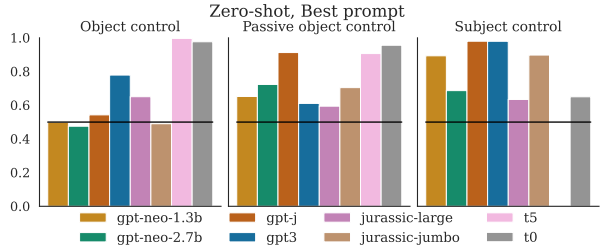


Figure 1: Zero-shot accuracy on object control, passive object control, and subject control. Black line represents random performance (50% accuracy).

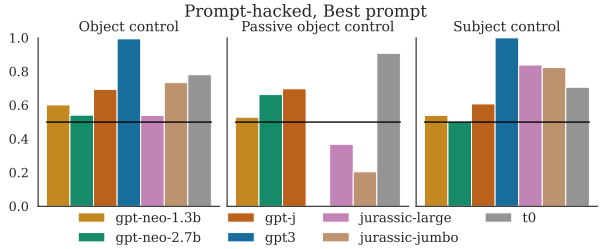


Figure 2: Accuracy on object control, passive object control, and subject control after prompting with agent and patient questions. Accuracy changes to Fig. 1 are generally consistent within heuristic groups.

The autoregressive models (parameters in Billions) considered are GPT-3 Davinci ($\sim$ 175B params), GPT-Neo (1.3B and 2.7B), GPT-J (6B), Jurassic Large (7.5B), and Jurassic Jumbo (178B). Details about training data can be found in Appendix C. The T2T models are: T5 finetuned for QA (220 million)[3] and T0pp (11B). 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.

**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. However, this method requires heuristics to classify the output strings into categories. In Appendix A.1 we validate our heuristics, verifying that for locally-run models the trends are similar when using logits associated with the correct answer. A full description of the heuristics is given in Appendix A. 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 Results and Analysis

In Fig. 1, we see that model classes have different results; we further classify models into 3 groups:

1. GPT-Neo and Jurassic Jumbo are better on subject and passive object control than object control. This is consistent with use of a positional heuristic, namely to take the *first* NP in the matrix clause (i.e. *Max* Distance Principle).[4]
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 stands alone, performing well on object control and nearly perfectly on subject control. However, poor performance on passive object control suggests a positional heuristic (take the second NP) being used for object control verbs, rather than an agency-based heuristic.

We make the following hypotheses for how prompting with questions about agents and patients will affect each group. For Group 1, if the models are following positional heuristics, the additional prompts will provide evidence inconsistent with the heuristic. This could lead to the heuristic being dropped; in that case, we would expect to see a drop in performance in passive object control and subject control (where the heuristic is beneficial) and an increase in performance in object control (where the heuristic is not beneficial). In Group 2, since the model outputs are already consistent with a semantic role-based explanation, we do not expect much change. However, in Group 3, we posit the existence of a heuristic for object control (to take the first NP) which has negative effects on performance for passive object control. Thus, as in Group 1, we expect that evidence against the heuristic in the form of prompts will boost performance in passive object control, while reducing performance on active object control.

Fig. 2 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 except Jurassic Jumbo in subject control. At the same time, all object control performance increases significantly

[3]Note that because of fine-tuned nature of the "T5 for QA" model, the expected prompt format is different (cf. Appendix D.2). Prompt hacking cannot be done on this version of T5, so it is only used in the first experiment.

[4]Jurassic Large is not included in this group since its performance is poor for all 3 settings.

for Group 1. These results confirm our hypotheses, supporting the notion that these models follow a positional rather than semantic heuristic. For Group 2 (T0), we find a slight but significant decrease in performance on both object control types, and no significant difference for subject control. Finally, for Group 3 (GPT-3) we 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. Note that the passive performance drop is from a lack of parseable strings being produced, rather than incorrect predictions.

**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 significantly worse on zero-shot object control examples (active and passive). 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 clarity on the training data used for GPT-3 and Jurassic Jumbo.

We observe also that larger models tend to have higher performance: GPT-J is significantly better on all settings than GPT-Neo-1.3B and 2.7B, and Jurassic Jumbo is significantly 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 last result suggests that perhaps larger models are more prone to following heuristics, even when they are wrong; however, additional evidence is needed.

**SPR labeling** 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. 1 is more consistent with an role-based heuristic. However, these are separate tasks – thus, it is possible for GPT-3 and GPT-J to follow non-role-based heuristics in one task while still encoding information about agent and patient properties. Finally, we note that in both Fig. 1 and Fig. 2, GPT-3 performs well on both subject control and object control in the active, which is consistent with it containing information about agency and patienthood.

| 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.

**Limitations** A major limitation is the use of fixed prompts: all models tested were found to be sensitive to the prompt format, and while a relatively large number of prompts were explored by varying instructions, names, verbs, and actions, it is possible that there are more optimal prompts for the task. In addition, the work is limited by the use of open generation. While open-ended generation allows for more flexibility than constrained decoding, it also introduces the challenge of interpreting the model outputs. We validate the use of open generation in Appendix A.1. We also note that both these limitations are also common in human subject research.

## 4 Related Work

Generally, the knowledge contained in such models has been measured with frozen models, in a probing setup using cloze-style prompts (Schick and Schütze, 2021). Large 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). With PLMs, this has often been done by recasting benchmark examples into text, either with zero examples (Sanh et al., 2021) or in the form of prompt hacking (Brown et al., 2020; Raffel et al., 2020). Ettinger (2020) present a suite of comparisons between PLMs and psycholinguistic experiments.

## 5 Conclusion

The results in Fig. 1 indicate that differences between models are not merely of degree, but of kind, with groups of models following wholly different strategies, some of which are inconsistent with the dominance of object control in English. This underscores the need for transparency in the reporting of model details, and especially of training data.

## References


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.

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.

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*.

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.

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.

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.

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.

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.

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*.

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.

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.

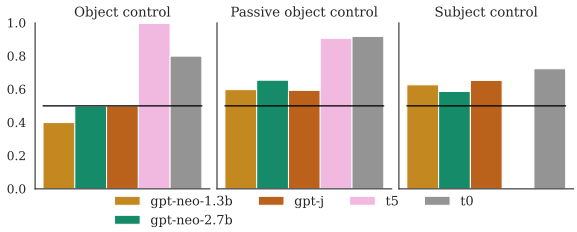


Figure 3: Accuracy of logit-scored model, taking the max across instruction types.

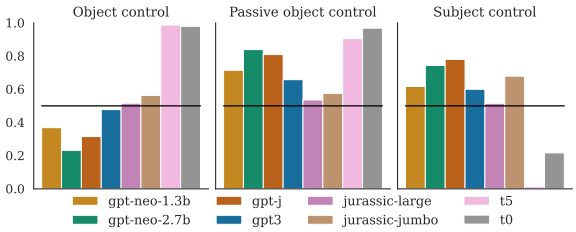


Figure 5: Accuracy of long instruction template on professions. Performance follows similar trends to comparable results with names (Fig. 4).

# A Metrics

The heuristics extract single word answers as well as answers like "The answer is: NAME". For some models and settings, the model re-generates the entire prompt before answering. Levenshtein distance is used 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 returns the first valid answer that is produced by the model.

## A.1 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 (all models run in Huggingface Transformers). Comparing the results to Fig. 1, we see similar but less pronounced trends. As before, GPT-based models perform better on subject control and passive object control, while T2T models perform better on object control and passive object control. This validates our choice to use heuristics rather than logits for the remaining results.

# B Profession Results

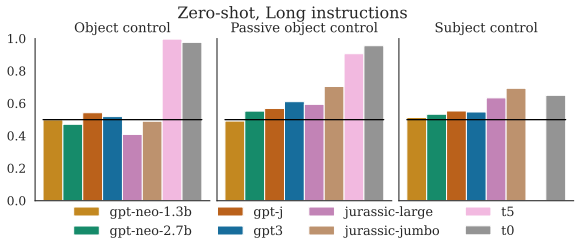


Figure 4: Accuracy of long instruction template on names.

# C Models

- **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.
- **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 text intended for pre-training. GPT-Neo has 3 sizes: 1.3B, 2.7B, and 6B parameters (GPT-J).
- **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 similar to GPT-3 Davinci, the details are unclear. Relevant differences to GPT-3 are in the tokenization (which includes multi-word expressions) and use of fewer, wider layers.

The T2T models we consider are:

- **T5 for QA**: A T5-base T2T model (220-million parameters) (Raffel et al., 2020) pre-trained on cleaned Common Crawl data (C4) and fine-tuned on SQuaD QA data (Rajpurkar et al., 2016).
- **T0pp**: presented by Sanh et al. (2021), T0PP is an 11B parameter T5 model pre-trained on C4, finetuned specifically for zero-shot QA on the P3 dataset of NLP benchmark data recast into prompts.

# D Data

## D.1 Frequency of Subject and Object Control

In Section 1, we claimed that object control is more frequent that subject control. To qualify this claim in the context of PLMs, 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, searching with templates similar to the

```
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 6: Zero-shot probe for object control. Colors indicate names, which are swapped.

```
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 7: A prompt-hacked example for object control, with long-form instructions.

sentences in examples 2 and 1. For object control, we use the same verb list as in Section 2. For subject control, we only use promise, as in Section 2. We sub-sample the first $1,000,000$ sentences of C4 and search it with the templates, finding that object control occurs $10,435$ times, while subject control occurs only $160$ times, i.e. object control is $\sim 65$ times more frequent.

### D.2 Control Prompts

An example prompt for zero-shot object control can be seen in Fig. 6. The colors indicate the names, which can be swapped out for other names. In this example, the long-form instructions are used and the order of the names has been swapped from the original order, which is the same order as the names in the context clause. In the short-form instructions, the phrase "Answer the question with either <name1> or <name2>" is removed, and the "<name1> or <name2>" clause is removed from the question line.

Fig. 7 shows an example prompt with prompt hacking to increase the salience of agents and patients. The additional questions do not provide any direct example of how to answer the test question, but they do identify the agent and patient in the matrix clause, raising the salience of the semantic roles.

### D.3 SPRL Prompts

Fig. 8 shows an example prompt for change of state, with a single prompting example preceding the test

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

example. For change of state, we experiment with two question formats. The first is shown above, the second asks, *"In the event* PRED*, does state of the participant* ARG *change?"*. For volition, the questions read: *"In the event* PRED*, does the participant* ARG *act with volition?"* and *"In the event* PRED*, does the participant* ARG *act on purpose?"*

### D.4 Licensing

All data and code is released under an MIT license.