# Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks

**Katherine M. Collins**[1,2*†], **Catherine Wong**[2*], **Jiahai Feng**[2],
**Megan Wei**[2], and **Joshua B. Tenenbaum**[2]
[1]University of Cambridge, [2]MIT
[†]kmc61@cam.ac.uk

## Abstract

Human language offers a powerful window into our thoughts – we tell stories, give explanations, and express our beliefs and goals through words. Abundant evidence also suggests that language plays a developmental role in structuring our learning. Here, we ask: how much of human-like *thinking* can be captured by learning statistical patterns in language alone? We first contribute a new challenge benchmark for comparing humans and distributional large language models (LLMs). Our benchmark contains two problem-solving domains (*planning* and *explanation* generation) and is designed to require generalization to new, out-of-distribution problems expressed in language. We find that humans are far more robust than LLMs on this benchmark. Next, we propose a hybrid *Parse-and-Solve* model, which augments distributional LLMs with a structured symbolic reasoning module. We find that this model shows more robust adaptation to out-of-distribution planning problems, demonstrating the promise of hybrid AI models for more human-like reasoning. **Keywords:** language; problem-solving; programs; language of thought; neuro-symbolic models

## Introduction

Language expresses the rich internal landscape of our thinking in a form that can be shared externally with others. We tell stories about real (*what did I do today?*) and hypothetical (*what would I do if I won the lottery?*) situations; give instructions for achieving goals ranging from the mundane (*how do I put away the dishes?*) to the complex (*how do I fix a carburetor?*); and propose explanations for both everyday events (*why isn't the light bulb turning on?*) and novel observations (*what's that strange beeping sound?*). Learning language and learning *from* language also play crucial roles in the development of children's thinking (Gopnik & Meltzoff, 1997; Carey, 2009; Harris et al., 2018). But what, in computational terms, is the relationship between language and thought, and between learning language and learning to think?

Classical theories draw a stark division between *thinking* as the manipulation of structured representations in an internal symbol system or language of thought (LOT) (Fodor, 1975), and *language* as a system of mappings between those representations and outwardly expressed forms (e.g., sounds, text). Under this view, learning language plays at best a supporting role in learning to think. Recently however, a new generation of statistical language learning systems in AI has put forth a serious challenge to this view. So-called *large language models* (LLMs) (Brown et al., 2020; Rae et al., 2021) have demonstrated such striking success in realistic language production that they often appear to be "thinking" – and yet they are driven solely by neural networks trained to predict the distribution of next words in long text sequences from very large corpora of human language. Other work has proposed using LLMs as a universal foundation for emulating many human reasoning abilities – including capacities as diverse as *physical reasoning* (Bisk et al., 2019), *task-level planning* (Sharma et al., 2021; Huang et al., 2022), and even *mathematical reasoning* (Cobbe et al., 2021) – simply by re-framing them as linguistic prediction. Under this view, "all you need is language": learning to think requires little more than learning (the statistics of) language, or learning only the latent structure sufficient to produce the most probable next word in any linguistic context.

In this paper, our goal is to critically assess how close modern LLMs come to actually learning to think, and to sketch out an alternative hybrid view of the language-thought interface that *integrates* elements of the classical LOT and recent LLM paradigms. In Part I, we describe a new, generic approach for constructing *linguistic reasoning prompts* that measure flexible, creative thinking abilities in novel situations, as opposed to the ability to retrieve familiar patterns of thought for familiar situations. We use an *iterative constraint generation* paradigm that extends initial linguistic prompts using linguistic *constraints* that restrict production of the most common human responses, forcing responses that require novel language production – and, we argue, a greater degree of thinking. We compare LLMs to humans using this benchmark on two domains – *plan* and *explanation* generation – and find that humans both significantly outperform LLMs in general, and are comparatively more robust to prompts that extend beyond the standard distribution of human language. In Part II, we propose an alternative computational approach that leverages an LLM to map natural language into a space

*Contributed equally.

‡Data and code for the project can be found at: https://github.com/collinskatie/structured_flexible_and_robust



In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*. 

of structured programs, such that reasoning problems can be solved by powerful, scalable symbolic algorithms - rather than the purely neural form of end-to-end LLMs alone. We implement and demonstrate this model in a simplified synthetic language setting designed to emulate the *planning domain* in Part I. Our results suggest that such hybrid approaches are a promising way forwards, albeit still rich with potential for future improvement.

## Part I: Linguistic reasoning benchmark for humans and language models

The first core motivation of this work is to evaluate the extent to which modeling the *predictive distribution of language* actually captures the underlying *reasoning* latent in human language. Towards this end, we propose a benchmark task (Fig. 1) based on two core reasoning abilities – *goal-based planning* and *causal explanation* – using an iterative design to challenge models which simply learn predictable responses from prior language.

### Methods

We benchmark human and language model performance using a two-stage experimental design. In the first stage, an iterative *human language production experiment* (Fig. 1B), we collect human responses on two domains (**planning** and **explanations**) under three progressively more challenging conditions: a baseline **initial prompt** condition using a collecting of linguistic reasoning prompts; and **two constrained conditions** which restrict the use of common answers to each prompt, in order to encourage participants to generate novel linguistic solutions. In the second stage, we evaluate a *large language model* (LLM) on the same prompts, and collect responses by sampling from its predictive distribution. We describe each stage in more detail below.

#### Human language production experiment

***Participants*** 240 participants recruited from Prolific (2 domains x 3 conditions x 40 participants) completed the task. Base pay was $15/hr, with a $1 quality bonus.

***Condition 1: initial reasoning prompts*** To measure baseline performance, our first reasoning condition elicits human responses to **initial prompts** (Fig. 1B, *Condition 1*) on each grounding domain. We construct 28 *goal prompts* for the planning domain (Fig. 1A, top), designed to elicit a concrete linguistic plan and to vary in their base typicality (eg. ranging from *clean the dirty dishes* to *get a sofa on the roof*). We also construct 28 *causal event* prompts of varying typicality for the explanations domain (Fig. 1A, bottom), inspired by the "unusual event" prompts in (Korman & Khemlani, 2020): each event begins with an inciting cause and its usual consequence, then poses a counterfactual.

Participants in this condition responded to a random batch (n=7) of prompts from a single domain, resulting in 10 unique responses per prompt. After responding to all prompts, we also ask participants to score *base typicality* for each prompt of the goal (on planning) or inciting event (on explanations) using a 7-point Likert scale.

***Condition 2 and 3: constrained reasoning prompts*** In the subsequent conditions (Fig. 1B, *Condition 2, 3*), we evaluate the human ability to flexibly generate more novel plans and explanations for the same initial prompts, by restricting their responses to prevent subjects from falling back on the most common solutions. Specifically, we use subject responses from Condition 1 to determine common (and likely highly *predictable*) components of plans and explanations for each prompt. We construct linguistic constraints by extracting concrete *nouns* from all responses to a given prompt (using an expert human tagger, who also lemmatizes and standarizes the form of each noun). We then extend each initial prompt in two more challenging conditions: in the **most common noun constrained** condition, we restrict responses which use the single most common noun; in the **all initial nouns constrained**, we restrict *all* nouns which appear in the initial responses.

A new set of participants responded to a random batch (n=7) of prompts in a single domain and condition, again resulting in 10 unique responses per prompt and condition that reflect these linguistic constraints.

**Language model matched production experiment** Our human experiment yields a series of linguistic prompts, in which individual goal and explanation prompts are extended across two more challenging conditions through linguistic *constraints* that restrict the usage of the most common responses to each.

We use these same prompts to construct a benchmark language production task for our artificial language model. We evaluate our prompts on the state-of-the-art model *GPT-3* (Brown et al., 2020), using the *few-shot prompting* technique introduced in (Brown et al., 2020) for generating predictive language for particular tasks. Specifically, we seed the model with a small number of *examples* (n=12 goals, and n=15 explanations: the maximum number of examples the model allowed, based on token limits) pairing heldout prompts and human-generated text, then elicit generated responses for each prompt across all conditions.

To eliminate purely degenerate text, we also *prescreen* the samples by asking human evaluators (N=370; recruited from Prolific) to score responses for surface language errors alone, and remove the lowest scoring responses. After screening, we collect a total of 20 LLM-generated responses for each prompt in each condition.

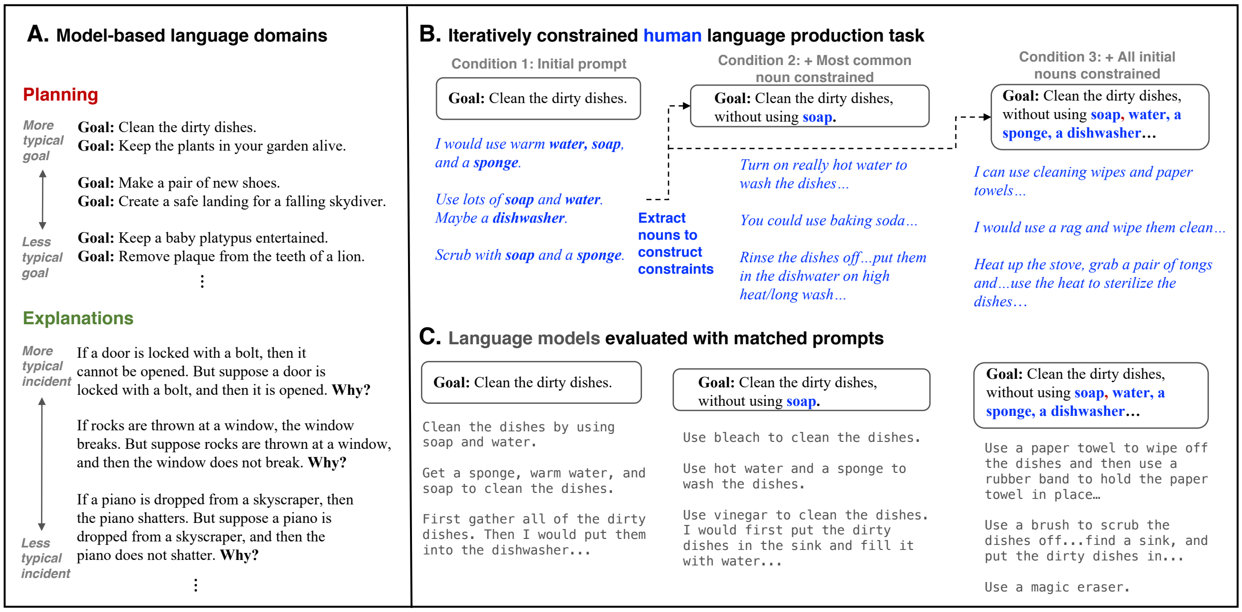


Figure 1: Iterative reasoning task overview. A) Sample goals and scenarios for the planning and explanation domains, respectively, illustrating the range of base typicality of our stimuli; B) Formation of constraints from human-generated language, where constraints are selected based on frequency, with sample human generations (blue text) C) LLM-generations (gray text) in response to the same prompts.

**Blind comparative human evaluation** Having collected human and LLM responses to the same linguistic prompts across all conditions, we now benchmark their relative performance using blind human evaluators (N=393; recruited from Prolific) asked to evaluate responses in a single domain and condition a 7-point Likert scale (1: worst; 7: best). Subjects rated responses for a random batch of *prompts*, scoring a (randomly shuffled) set of human (n=10) and LLM (n=10) responses for each.

## Results

Representative *human* responses and *language model* responses across both domains and conditions are depicted in Fig. 2. To investigate comparative performance, we fit linear mixed effects regression (LMER) models predicting the human-evaluated score and use a corresponding likelihood ratio test (LRT) between an ablated model to determine the significance of the fixed effects. Fig. 3 shows results of the blind human evaluation, and depicts statistical significance within and across conditions.

**People outperform the LLM within each reasoning condition** We first fit a LMER predicting the human evaluated score from the *source* language generator (human or LLM), with random effects for the individual raters and prompts (syntax: `score ~ source + (1 | rater_id) + (1 | prompt)`). Our LRT finds that there is a significant effect ($p < 0.001$) of the language source (humans vs. LLM) in both domains and in each condition (3, black indicators), humans outperform the LLM in *every condition*, across both domains.

**People are more robust to out-of-distribution prompts with constraints** We next consider our more central question: how well do language models perform specifically on our more **constrained** conditions, designed explicitly to force both humans and models to generate novel solutions to our underlying reasoning task? We expect humans to not only outperform language models in a direct comparison across individual prompts, but also to be comparatively more *robust* to prompts which restrict highly predictable answers, and require responses beyond the distribution of standard human language.

An initial LMER with a fixed effect for the *condition* (*unconstrained*, *most common constraint*, or *many constraints*) suggests that both humans and LLMs are sensitive to the added constraints, though we find a strongly significant effect of condition on performance for LLMs ($p < 0.001$); and a weakly significant effect ($p = 0.03$) for humans in the planning domain but strongly significant for explanations ($p < 0.001$).

However, a subsequent LMER with an interaction term for the language source (humans or LLMs) and condition (fit pairwise across each successive set of conditions) indicates that humans and LLMs are not *equally* sensitive to constraints: we find strongly signficant interaction terms

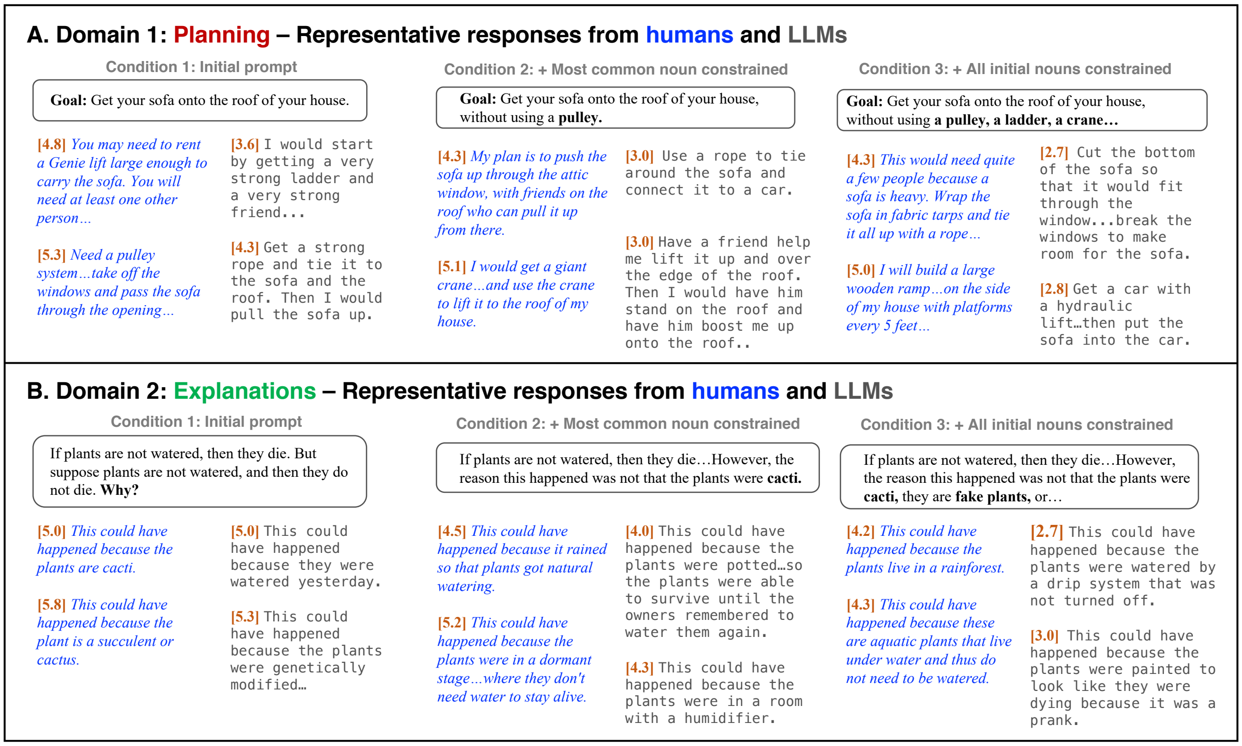


Figure 2: Representative plans (A) and explanations (B), per constraint condition, generated by humans and an end-to-end LLM. Average goodness rating, over the human evaluators for each generation, is shown in orange.

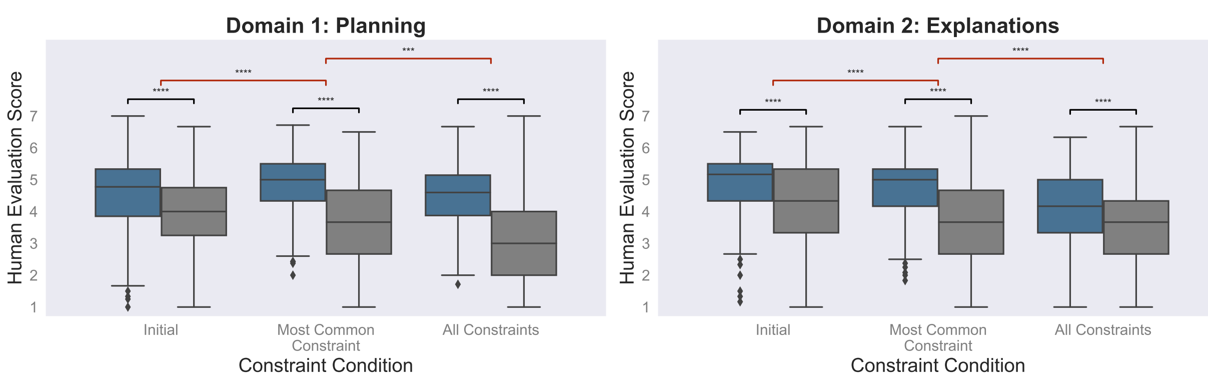


Figure 3: Mean overall goodness rating over plans (left) and explanations (right), show across all three constraint conditions. Humans (blue boxes) significantly outperform the LLM (gray boxes) in every condition (black, lower bars) and in successive pairwise conditions (red, upper bars).

(Fig. 3, red) indicating that humans are more *robust* to added constraints across each condition. This supports our central hypothesis: language models are increasingly poor at solving the underlying task once the prompts are constrained to restrict predictable responses.

**People are more robust to *goal* typicality** We also investigate whether another measure of linguistic predictability – the atypicality of our base prompts – also impacts LLM performance relative to humans. We fit a final LMER model with an interaction term for source and human typicality scores elicited in our initial experiment. Interestingly, we find a significant interaction effect of typicality ($p < 0.001$) for the *planning* domain, but not for explanations. As assessing typicality for these prompts is more complex, further work (such as linguistic measures of prompt typicality) are necessary to better assess the explanations domains. This finding further supports our broader hypothesis: that LLMs are less robust to responding to out-of-distribution scenarios which pose novel, but solvable, planning problems.

**Qualitative analysis of commonsense failures in LLM reasoning** Do large language models suffer from distinctively *different* patterns of errors? An initial, qualitative examination suggests that large language models are particularly prone to errors indicating a more fundamental lack of "common sense" understanding: of the underlying task, or the world knowledge required to solve it. A preliminary examination suggests that language models struggle particularly in generating coherent, realistic solutions for problems that require novel but concrete physical reasoning: as in the *sofa on a roof* goals in Fig. 2; or failures to understand *color* (*The carpet was white, so the blue dye did not show up*); *water* (the *grass is not made of water and so it does not absorb the water*); or gravity and *material* (eg. someone failing to scrape their knees after falling in *pants that were made of paper*). Taken together, our reasoning experiment suggests that despite the surface plausibilty of their generated text, large language models generally struggle to emulate the latent reasoning that backs human responses – once problems expressed in language require solutions beyond the standard, and most predictable, distribution of prior language, the apparent "reasoning" abilities of these models deteriorate sharply.

## Part II: Integrating language with structured reasoning models

Our results in Part I suggest that even very large language models may not capture the characteristic flexibility of human reasoning: they struggle to produce language reflecting novel computation over an underlying task.

Here, we propose an alternate computational approach for reasoning about problems posed in language. Rather than hoping to simulate latent computations (like planning) by directly predicting output language, we propose a simple (but demonstrative) *parse-and-symbolic planner* (P+S) model which grounds language in an explicit "language-of-thought" (Fodor, 1975): a formal *program* expressing the meaning of the linguistic prompt, which interfaces with a symbolic computational solver (Fig. 4B).

### Simulated planning experiment

We introduce a *simulated planning domain* to benchmark our *parse-and-symbolic planner* model against a standard LLM (here, GPT-Neo (Black, Gao, Wang, Leahy, & Biderman, 2021)), using a restricted set of prompts designed to emulate the core properties of the broader planning domain in Part I. We focus on *planning* here for a straightforward metric of comparative performance: accuracy of our restricted plans can be evaluated directly on an explicit world model.

**Initial and constrained synthetic planning prompts** As with Part I, our simulated experiment benchmarks model performance under three progressively more challenging conditions: responses to an **initial set** of linguistic goal prompts (Fig. 4A, *Condition 1*); and **two constrained conditions** which introduce new linguistic constraints over the initial goal (Fig. 4B, *Condition 2, 3*). As is obvious from Fig. 4B, our conditions differ from Part I in one important respect: we extend our initial goals with *positive constraints*, rather than the negative constraints in Part I. This format permits a more direct, albeit simplified, evaluation of the core task – fully simulating *restrictions* on initial resources would require modeling (and communicating) all possible alternative ways to achieve a goal in a simulated environment – while still requiring models to reason about complex, out-of-distribution language.

We generate initial and constrained goal prompts – along with a linguistic *initial condition* completely specifying the starting planning state for each prompt – from a synthetic grammar over a simple *object-stacking domain* (Gupta & Nau, 1992), in which each goal is a target stack of objects on a table (Fig. 4). *Initial prompts* involve goals with a single common household object; these are extended with both a **single constraint** and **many constraints** (n=4) that introduce additional, *unusual* objects into the initial goal. In total, we sample n=100 initial goals and then sample constraints for both extended conditions.

**Parse-and-solve model** Fig. 4B depicts a schematic of our parse-and-solve model, designed to disentangle language from the underlying computation required to solve planning tasks expressed in language. Our model integrates two distinct components. First, it *parses* language into a formal *program* representing the initial problem state and goal (using the PDDL planning language (McDermott et al., 1998)). For more direct comparison with a benchmark LLM, we also use a *large language model* as our surface *parser*: we use the Codex (Chen et al., 2021) model (a GPT-3 model fine-tuned on a joint distribution of language and symbolic programs), which can "parse" language into programs using an analogous few-shot prompting technique (seeded with coupled examples of text and code). Unlike our comparison model, however, we employ distributional prediction only for a more constrained task: emulating the joint variation between a natural and formal language. The parsed *programs* are passed to our model's second core component: a symbolic *solver*, modeled with a search-based planner (Alkhazraji et al., 2020) which attempts to generate a symbolic plan over a restricted set of actions (*moving* objects from one location to the next) to solve the parsed goal.

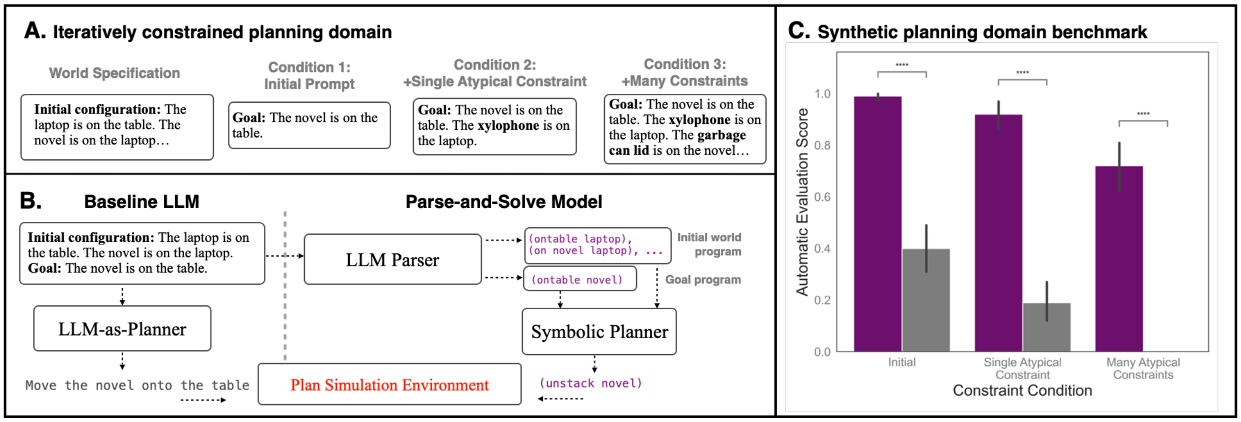


Figure 4: Simulated iterative planning task overview. A) Example progressively-constrained goal stimuli; B) Evaluation compares plans generated directly from an LLM (left) with plans generated from P+S (right); C) Success rate of P+S model (purple) vs. LLM (gray); P+S statistically significantly outperforms the LLM under each condition (black bars).

**Plan simulation environment** Unlike in Part I, plans using the restricted space of actions in this domain can be simulated directly to assess accuracy. The P+S model outputs executable PDDL actions; the LLM-as-planner baseline outputs language which we reparse by inverting the synthetic grammar into PDDL actions. For both models, we mark unparseable or invalid plans as unsuccessful.

## Results

Analogous analyses to those in Part I (Fig. 4C) – measuring the comparative performance of our model with an LLM, as well as its *robustness* to constraints – suggest that our hybrid model, which uses predictive modeling only to transform language into a structured interface to an underlying symbolic planner, vastly improves its ability to adapt to complexly constrained goals.

**Parse-and-solve model outperforms LLM** An LMER comparing our two models (P+S and LLM) finds a strongly significant difference in overall performance ($p < 0.001$; Fig. 4C): indeed, the LLM solves *none* of the problems in our most constrained condition.

**Comparative robustness to constraints** Interestingly, a pairwise LMER testing for an interaction between source and condition does not find a significant interaction effect, suggesting that both models decline similarly in relative performance between conditions. One likely possibility is that this is an artifact of our restricted experiment size: the LLM simply can perform no worse in the final condition. However, these results could also suggest that the parsing approach we use here – which employs distributional models to map language into programs – may itself struggle to generalize; a hybrid *parser*, which itself draws on more structured representations (like classical *linguistic* grammars), might be better suited to parsing our most challenging compositional goals.

## Discussion

Human language provides a richly structured window into how we think about the world. Our results, however, suggest that modeling the distribution of language alone may not be sufficient to capture the computations underlying planning, explanations, and other forms of reasoning which ground the language we produce. Instead, we propose an alternative approach: hybrid models which use distributional prediction to map language into structured formal representations of meaning that interface directly with structured symbolic algorithms (Ellis et al., 2020; Wong et al., 2021; Nye et al., 2021). Our contributions here leave much open for future work: to more systematically characterize regimes under which simply producing probable language closely approximates, and *deviates*, from human reasoning, and go beyond the simple demonstration model we have provided towards broader-coverage models for more realistic reasoning domains.

An important next step will be building on the qualitative analyses in Part I to disentangle the many factors (e.g., *accuracy*, *semantic coherence*, and *concision*) that may separate human performance from purely predictive responses. In tandem, the hybrid model we propose here offers a promising, albeit highly restricted, step towards emulating human-like reasoning over language. How do we *learn* the structured world models, or even sophisticated planning algorithms, that our simple model builds upon? Our core modeling approach suggests a path towards these more fundamental learning problems: using language to construct, or guide discovery, of *programs* which represent novel environments, actions, and even algorithms for operating over such worlds.

## Acknowledgments

We thank Laura Schulz, Junyi Chu, Alex Lew, Joao Loula Guimaraes de Campos, Max Nye, and the rest of the GPS Community for many thrilling, inspiring conversations, as well as practical advice with our project. We are also deeply grateful for Pratyusha Sharma, Jacob Andreas, and Noa Korneev for their thoughtful suggestions and support. We thank Yoni Friedman for his fantastic help with human annotator recruitment. Additionally, we thank the OpenAI team for increasing our quota to enable us to run more GPT-3 rollouts for Part I, and our Anonymous Reviewers for useful comments.

KMC is supported by a Marshall Scholarship and conducted work on the project under a Goldwater Scholarship. CW and JBT are supported by AFOSRFA9550-19-1-0269, the MIT Quest for Intelligence, the MIT-IBM Watson AI Lab, ONR Science of AI, and DARPA Machine Common Sense.

## References

Alkhazraji, Y., Frorath, M., Grützner, M., Helmert, M., Liebetraut, T., Mattmüller, R., … Wülfing, J. (2020). *Pyperplan.* https://doi.org/10.5281/zenodo.3700819. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.3700819 doi: 10.5281/zenodo.3700819

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., & Choi, Y. (2019). PIQA: reasoning about physical commonsense in natural language. *CoRR*, *abs/1911.11641*. Retrieved from http://arxiv.org/abs/1911.11641

Black, S., Gao, L., Wang, P., Leahy, C., & Biderman, S. (2021, March). *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.5297715 (If you use this software, please cite it using these metadata.) doi: 10.5281/zenodo.5297715

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *CoRR*, *abs/2005.14165*. Retrieved from https://arxiv.org/abs/2005.14165

Carey, S. (2009). Where our number concepts come from. *The Journal of philosophy*, *106*(4), 220.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., … Zaremba, W. (2021). Evaluating large language models trained on code. *CoRR*, *abs/2107.03374*. Retrieved from https://arxiv.org/abs/2107.03374

Cobbe, K., Kosaraju, V., Bavarian, M., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ellis, K., Wong, C., Nye, M. I., Sablé-Meyer, M., Cary, L., Morales, L., … Tenenbaum, J. B. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *CoRR*, *abs/2006.08381*. Retrieved from https://arxiv.org/abs/2006.08381

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.

Gupta, N., & Nau, D. S. (1992). On the complexity of blocks-world planning. *Artificial Intelligence*, *56*(2-3), 223–254.

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, *69*, 251–273.

Huang, W., Abbeel, P., Pathak, D., & Mordatch, I. (2022). *Language models as zero-shot planners: Extracting actionable knowledge for embodied agents.*

Korman, J., & Khemlani, S. (2020). Explanatory completeness. *Acta Psychologica*, *209*, 103139. Retrieved from https://www.sciencedirect.com/science/article/pii/S0001691819303531 doi: https://doi.org/10.1016/j.actpsy.2020.103139

McDermott, D., Ghallab, M., Howe, A., Knoblock, C., Ram, A., Veloso, M., … Wilkins, D. (1998). *Pddl - the planning domain definition language* (Tech. Rep. No. TR-98-003). Yale Center for Computational Vision and Control,.

Nye, M. I., Tessler, M. H., Tenenbaum, J. B., & Lake, B. M. (2021). Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *CoRR*, *abs/2107.02794*. Retrieved from https://arxiv.org/abs/2107.02794

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., … others (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Sharma, P., Torralba, A., & Andreas, J. (2021). *Skill induction and planning with latent language.*

Wong, C., Ellis, K., Tenenbaum, J. B., & Andreas, J. (2021). Leveraging language to learn program abstractions and search heuristics. *CoRR*, *abs/2106.11053*. Retrieved from https://arxiv.org/abs/2106.11053