Language Models are Crossword Solvers

Soumadeep Saha, Sutanoya Chakraborty, Saptarshi Saha, Utpal Garain
Indian Statistical Institute, Kolkata

CONTACT

Objectives

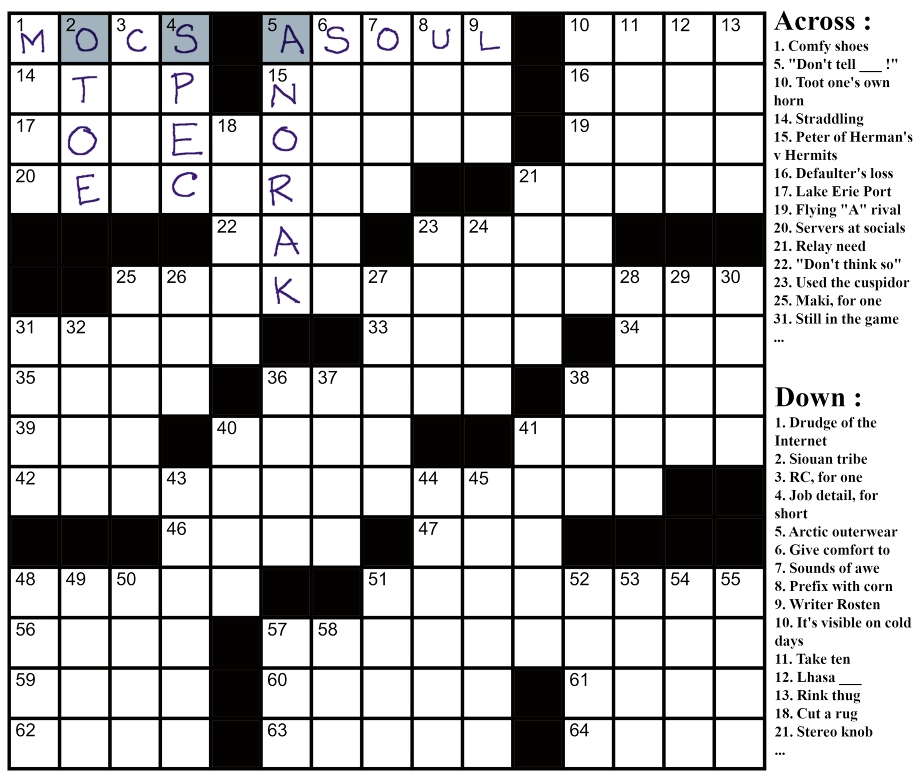

Figure: **Example of a crossword puzzle.**

- Constrained language generation with LLMs.
- Crosswords are a type of constrained word puzzle requiring proficiency in understanding contextual clues, semantics, wordplay, character manipulation, arithmetic, world-knowledge, multi-hop reasoning, etc. (see Fig. 1, 2).
- Analyze LLMs' ability at this task with the **primary goal of understanding strengths and weaknesses demonstrated by SoTA LLMs.**

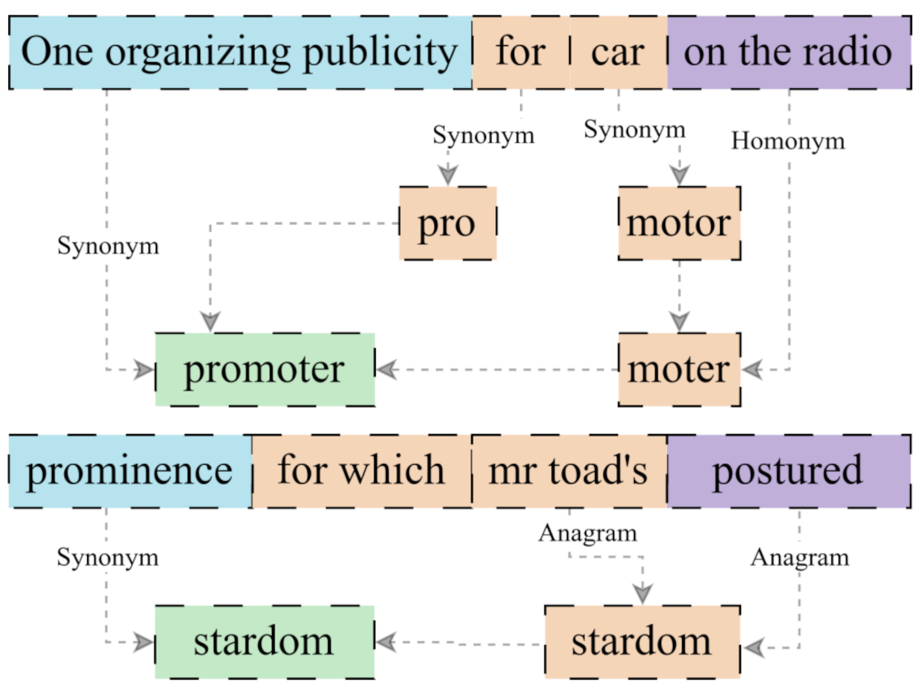

Figure: **Examples of cryptic crossword clues.**

Background

There are specialized *straight* crossword puzzle solving systems, reliant on large clue-answer databases and CSP algorithms [4, 7]. Our aim is not creating a specialized crossword solver, but employing LLMs for constrained generation. Solving *cryptic* crosswords with large clue datasets and a CFG parser [1] has shown poor performance, as has training small LMs (T5) [2, 5, 6]. [3] attempted to solve NYT crossword puzzles with LMs and an SMT solver with limited success.

Experiments

Clue solving task - LM is given the clue and the length of the answer. The models demonstrate improved performance with scale across datasets and, show remarkable improvement on the NYT dataset with Llama 3 70B, GPT 3.5 Turbo, Claude 3 Sonnet, and GPT-4-Turbo achieving 27.2%, 26.05%, 37.7% and **41.2%** accuracy (EM), respectively (Fig. 3).

Hinted clue solving task - LLMs can successfully exploit constraints (letter patterns) to improve performance (Tab. 1) $\Rightarrow$ they might be able to solve full crosswords.

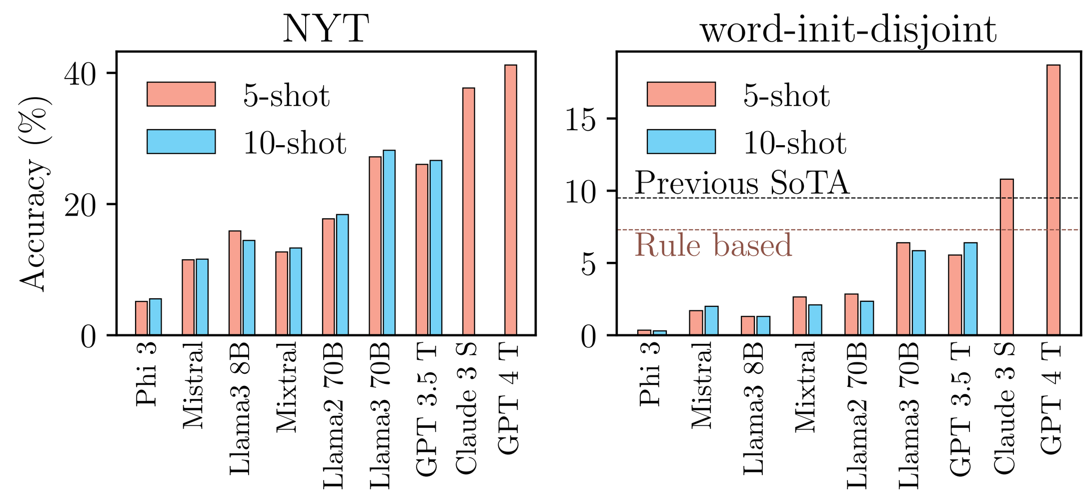

Figure: **SoTA LLMs with 5-shot prompts can answer crossword clues.**

Hint (%)	0%		25%		50%	
	NYT	*init*	NYT	*init*	NYT	*init*
Mistral 7B	10.95%	1.70%	9.70%	2.80%	11.95%	4.80%
LlaMa 3 8B	15.8%	1.30%	19.7%	2.85%	24.65%	6.25%
LlaMa 3 70B	27.20%	6.40%	31.80%	11.45%	45.30%	20.35%
GPT 4 Turbo	41.2%	18.70%	59.95%	33.70%	75.75%	52.85%

Table: **LLMs can improve by exploiting character constraints**. [6] reported 27.0% accuracy (70% hinted clues, fine-tuned Mistral). *GPT-4-Turbo* (**76.30%** accuracy) *outperforms it by a factor of* $\sim$**2.8**$\times$ *without fine-tuning*.

SweepClip Algorithm

- We address the problem of filling crossword grids with LLM assistance.
- This task involves constraint satisfaction in addition to answer generation.
- Our algorithm first generates a set of candidate answers for all clues (*sweep*) and uses a graph-based criterion (largest-connected component) to eliminate answers that do not fit (*clip*).
- Following this, we use the constraints from the previous step to generate more candidate answers.

Error Tolerance	% of Crosswords	
	LLaMa 3	GPT-4 T
100% solved	0	**48**
$\leq$ 1 character error	1	**55**
$\leq$ 5 character error	10	**71**
$\geq$ 90% Accuracy	30	**80**
$\geq$ 50% Accuracy	82	**98**

Table: **Results from solving NYT crosswords with our algorithm *SweepClip*.**

Generalizability & Reasoning

Model	*Guardian*	*init*
Llama 3 70B	5.5 %	6.4 %
Claude 3 Sonnet	12.5%	10.8%
GPT 4 Turbo	18.5%	18.7%

Table: **No performance dip on post-cutoff dataset.**

We see no appreciable difference in performance on the post-cutoff dataset (see Tab. 3), suggesting that LLMs can generalize beyond potential contamination.

Human evaluation was performed to assess reasoning ability with cryptic crossword clues (3-shot CoT prompt + GPT-4-Turbo). We found that **74%** of the time GPT-4-Turbo provided a correct answer, it also gave **sound reasoning** (no logical or factual errors) in support of the answer.

Sub-token Counting

SoTA LLMs struggle with adherence to length constraints, i.e., they show an inability to count characters within words or phrases (*sub-token counting*). If LMs could count, we should see no difference in performance across words with different prevalence, however, we find a significant difference in counting accuracy between *vocabulary* and *gibberish* words (Tab. 4).

Model	Vocab. (%)	Gibberish (%)
Phi 3 3.8B Instruct	79.4	61.2
Mistral 7B Instruct	47.9	28.2
Llama 3 8B Instruct	92.6	69.7
Mixtral 8x7B	92.6	80.1
Llama 2 70B	92.8	80.0
Llama 3 70B	99.6	87.5
GPT 3.5 Turbo	86.0	62.1
GPT 4 Turbo	99.8	98.8

Table: **LLM counting accuracy is affected by prevalence of words.**

To measure the effect of *sub-token counting performance* on clue-solving, we consider all such clues for which the model correctly deduced the semantics of the clue but failed to adhere to the length constraints (e.g., LECTURER $\leftrightarrow$ PROFESSOR, NANNA $\leftrightarrow$ GRANNY, etc.). GPT-4-Turbo and Llama 3 70B produce predictions with length errors **46.4%** and **59.9%** of the time, respectively, suggesting that this is a major roadblock.

Conclusions

- Constrained language generation is an increasingly relevant problem, and crosswords are a great benchmark in this regard.
- SoTA LLMs demonstrate the ability to solve crossword clues, and exploit constraints from partially solved grids.
- This ability generalizes to post-cutoff datasets, and sound reasoning is often produced in support of answers.
- Our algorithm *SweepClip* can solve (straight) crosswords with the aid of LLMs. This is the first successful demonstration of crossword solving with an out-of-the-box foundational LLM.
- LLMs' inability to count and adhere to length constraints is a major hurdle requiring further investigation.

References

[1] Deits, R. *github - rdeits/cryptics*, 2015.
[2] Efrat, A., et al. *Cryptonite: A cryptic crossword benchmark for extreme ambiguity in language*. In *EMNLP 2021*. 2021. doi:10.18653/v1/2021.emnlp-main.344.
[3] Kulshreshtha, S., et al. *Down and across: Introducing crossword-solving as a new NLP benchmark*. In *ACL 2022 (long)*. 2022. doi:10.18653/v1/2022.acl-long.189.
[4] Littman, M. L., et al. *A probabilistic approach to solving crossword puzzles*. *Artificial Intelligence*, 134(1):23–55, 2002. doi:https://doi.org/10.1016/S0004-3702(01)00114-X.
[5] Rozner, J., et al. *Decrypting cryptic crosswords: Semantically complex wordplay puzzles as a target for nlp*. In *NeurIPS 2021*, vol. 34, 11409–11421. 2021.
[6] Sadallah, A. B., et al. *Are llms good cryptic crossword solvers?*, 2024.
[7] Wallace, E., et al. *Automated crossword solving*. In *ACL 2022 (long)*. 2022. doi:10.18653/v1/2022.acl-long.219.

www.arxiv.org/2406.09043
espressovi.github.io