CrosswordLLM: A Benchmark Framework for Evaluating Large Language Models on Cryptic Linguistic Reasoning in Crossword Puzzles

Alex Frugé

April 16, 2025

Abstract

I'll fill this out later once I get some results, and finalize any fine-tuning steps I plan on making (embeddings, prompt engineering, etc).

Introduction

Large Language Models (LLMs) have demonstrated remarkable proficiency in natural language understanding and generation. However, their ability to solve complex linguistic tasks, such as crosswords, remains an open area of investigation. This project aims to design an evaluation pipeline to test how well different LLMs can solve crosswords. By analyzing performance variations across model sizes and architectures, I seek to understand whether larger models inherently perform better, or if architectural differences contribute more significantly to success. Additionally, I will explore how contextual information, prompt engineering, and multimodal inputs may affect model performance. This research will not only offer insights into the cognitive capabilities of LLMs but also contribute to broader AI evaluations in natural language processing (NLP).

Data Overview

Sourcing & Explanation

The data for this project was pulled from two different sources. The first source is Darin Hawley's New York Times Crossword Clues & Answers 1993-2021 dataset. [**Hawley_2021**] This dataset contains 781,573 entries, each of which come with the date the puzzle containing the clue was published, the clue itself, and the answer.

The NYT dataset is being used as our training data, as it contains a good spread of clue variety, including some pop-culture related clues, some wordplay, and a wide variety of difficulty in the clues themselves. It also contains a good number of clues involving multiple words in a clue being concatenated into one continuous word (see Table 1 for some examples of this). I'll also use the NYT dataset for a testing suite that tests the selected models on more conventional clues.

Date	Answer	Clue
11/8/2013	EATAT	Irritate
7/8/2012	ALQAEDA	War on terror target
7/7/2012	ASGOODASGOLD	100% reliable

Table 1: *Example clues from NYT Crossword Clues.*

The second source of clues comes from George Ho's Cryptic Crossword Clues dataset [1]. Cryptic crosswords are crossword puzzles that are more popular in the United Kingdom, and they are known to be highly difficult to solve. There are many ways to construct a cryptic crossword clue, such as using anagrams, hidden words, homophones, or a combination of wordplay techniques. **[Maynes-Aminzade_Henriq_2019]**

An example of these two elements coming together would be for the clue "Record is set in Washington", with the corresponding clue "DISC", considered as "IS" inside of "DC", which can be ascertained from the clue (record is another word for disc, and "IS" being literally "set" within Washington ("DC" for short)).

The goal with the cryptic clues dataset is to create a test for the trained models to see how well they can handle wordplay and deeper reasoning with language. I'd be pleasantly surprised to see the models handle these clues well, as it takes humans a long time to process these answers as well.

1. Donec dolor arcu, rutrum id molestie in, viverra sed diam
2. Curabitur feugiat
3. Turpis sed auctor facilisis

Mauris interdum porttitor fringilla. Proin tincidunt sodales leo at ornare. Donec tempus magna non

[1] George Ho's Cryptic Crossword Clues

mauris gravida luctus. Cras vitae arcu vitae mauris eleifend scelerisque. Nam sem sapien, vulputate nec felis eu, blandit convallis risus. Pellentesque sollicitudin venenatis tincidunt. In et ipsum libero. Nullam tempor ligula a massa convallis pellentesque.

This is a numbered list:

Species Identification

Proin lobortis efficitur dictum. Pellentesque vitae pharetra eros, quis dignissim magna. Sed tellus leo, semper non vestibulum vel, tincidunt eu mi. Aenean pretium ut velit sed facilisis. Ut placerat urna facilisis dolor suscipit vehicula. Ut ut auctor nunc. Nulla non massa eros. Proin rhoncus arcu odio, eu lobortis metus sollicitudin eu. Duis maximus ex dui, id bibendum diam dignissim id. Aliquam quis lorem lorem. Phasellus sagittis aliquet dolor, vulputate cursus dolor convallis vel. Suspendisse eu tellus feugiat, bibendum lectus quis, fermentum nunc. Nunc euismod condimentum magna nec bibendum. Curabitur elementum nibh eu sem cursus, eu aliquam leo rutrum. Sed bibendum augue sit amet pharetra ullamcorper. Aenean congue sit amet tortor vitae feugiat.

Mauris interdum porttitor fringilla. Proin tincidunt sodales leo at ornare. Donec tempus magna non mauris gravida luctus. Cras vitae arcu vitae mauris eleifend scelerisque. Nam sem sapien, vulputate nec felis eu, blandit convallis risus. Pellentesque sollicitudin venenatis tincidunt. In et ipsum libero. Nullam tempor ligula a massa convallis pellentesque.

Data Analysis

Vestibulum sodales orci a nisi interdum tristique. In dictum vehicula dui, eget bibendum purus elementum eu. Pellentesque lobortis mattis mauris, non feugiat dolor vulputate a. Cras porttitor dapibus lacus at pulvinar. Praesent eu nunc et libero porttitor malesuada tempus quis massa. Aenean cursus ipsum a velit ultricies sagittis. Sed non leo ullamcorper, suscipit massa ut, pulvinar erat. Aliquam erat volutpat. Nulla non lacus vitae mi placerat tincidunt et ac diam. Aliquam tincidunt augue sem, ut vestibulum est volutpat eget. Suspendisse potenti. Integer condimentum, risus nec maximus elementum, lacus purus porta arcu, at ultrices diam nisl eget urna. Curabitur sollicitudin diam quis sollicitudin varius. Ut porta erat ornare laoreet euismod. In tincidunt purus dui, nec egestas dui convallis non. In vestibulum ipsum in dictum scelerisque.

Mauris interdum porttitor fringilla. Proin tincidunt sodales leo at ornare. Donec tempus magna non mauris gravida luctus. Cras vitae arcu vitae mauris eleifend scelerisque. Nam sem sapien, vulputate nec felis eu, blandit convallis risus. Pellentesque sollicitudin venenatis tincidunt. In et ipsum libero. Nullam tempor ligula a massa convallis pellentesque. Mauris interdum porttitor fringilla. Proin tincidunt sodales leo at ornare. Donec tempus magna non mauris gravida luctus. Cras vitae arcu vitae mauris eleifend scelerisque. Nam sem sapien, vulputate nec felis eu, blandit convallis risus. Pellentesque sollicitudin venenatis tincidunt. In et ipsum libero. Nullam tempor ligula a massa convallis pellentesque.

Model	Training Time (s)	Loss (Epoch 5)
pythia-14m	444.85	0.8150
pythia-31m	635.19	0.6695
pythia-70m	1,101.80	0.5146
pythia-160m	2,594.33	0.3911
pythia-410m	7,197.65	0.2740
pythia-1.3b	XXX	YYY

Table 2: *First run across Pythia Suite using NYT Data (n=12,800, batch size=32, epochs=5)*

Results

Referencing a table using its label: Table **??**.

Aenean feugiat pellentesque venenatis. Sed faucibus tristique tortor vel ultrices. Donec consequat tellus sapien. Nam bibendum urna mauris, eget sagittis justo gravida vel. Mauris nisi lacus, malesuada sit amet neque ut, venenatis tempor orci. Curabitur feugiat sagittis molestie. Duis euismod arcu vitae quam scelerisque facilisis. Praesent volutpat eleifend tortor, in malesuada dui egestas id. Donec finibus ac risus sed pellentesque. Donec malesuada non magna nec feugiat. Mauris eget nibh nec orci congue porttitor vitae eu erat. Sed commodo ipsum ipsum, in elementum neque gravida euismod. Cras mi lacus, pulvinar ut sapien ut, rutrum sagittis dui. Donec non est a metus varius finibus. Pellentesque rutrum pellentesque ligula, vitae accumsan nulla hendrerit ut.

Aenean porttitor eros non pharetra congue. Proin in odio in dolor luctus auctor ac et mi. Etiam euismod mi sed lectus fringilla pretium. Phasellus tristique maximus lectus et sodales. Mauris feugiat ligula quis semper luctus. Nam sit amet felis sed leo fermentum aliquet. Mauris arcu dui, posuere id sem eget, cursus pulvinar mi. Donec nec lacus non lectus fermentum scelerisque et at nibh. Sed tristique, metus ac vestibulum porta, tortor lectus placerat lorem, et convallis tellus dolor eget ante. Pellentesque dui ligula, hendrerit a purus et, volutpat tempor lectus. Mauris nec purus nec mauris rhoncus pellentesque. Quisque quis diam sed est lacinia congue. Donec magna est,

Table 3: *Example two column table with fixed-width columns.*

Location		
East Distance	West Distance	Count
100km	200km	422
350km	1000km	1833
600km	1200km	890

hendrerit sed metus vel, accumsan rutrum nibh.

Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Etiam cursus lectus purus, tempus iaculis quam dictum tristique. Nam interdum sapien nec tempor mattis. Quisque id sapien nisi. Mauris vehicula ornare eros vel efficitur. Nulla consectetur, turpis quis fringilla tincidunt, mi neque iaculis lectus, vel commodo elit odio non ex. Duis facilisis, purus ac viverra iaculis, turpis lectus ultrices ante, ac vestibulum ligula magna in libero. Etiam tristique maximus lacinia. Vestibulum hendrerit, lacus malesuada laoreet blandit, sapien velit sollicitudin nunc, eu porttitor urna ligula at lorem. Aliquam faucibus eros in fermentum venenatis. Fusce consectetur congue pellentesque. Suspendisse at nisi sit amet est porttitor cursus. Cras placerat faucibus nunc, a laoreet justo dignissim sit amet.

International Support

àáâäãåèéêëìíîïòóôöõøùúûüÿýñçčšž
ÀÁÂÄÃÅÈÉÊËÌÍÎÏÒÓÔÖÕØÙÚÛÜŸÝÑ
ßÇŒÆČŠŽ

Links

This is a clickable URL link: LaTeX Templates. This is a clickable email link: vel@latextemplates.com. This is a clickable monospaced URL link: `https://www.LaTeXTemplates.com`.

Discussion

This statement requires citation [**Smith:2023qr**]. This statement requires multiple citations [**Smith:2023qr**, **Smith:2024jd**]. This statement contains an in-text citation, for directly referring to a citation like so: **Smith:2024jd**.

Subsection One

Suspendisse potenti. Vivamus suscipit dapibus metus. Proin auctor iaculis ex, id fermentum lectus dapibus tristique. Nullam maximus eros eget leo pretium dapibus. Nunc in auctor erat, id interdum risus. Suspendisse aliquet vehicula accumsan. In vestibulum efficitur dictum. Sed ultrices, libero nec fringilla feugiat, elit massa auctor ligula, vehicula tempor ligula felis in lectus. Suspendisse sem dui, pharetra ut sodales eu, suscipit sit amet felis. Donec pretium viverra ante, ac pulvinar eros. Suspendisse gravida consectetur urna. Pellentesque vitae leo porta, imperdiet eros eget, posuere sem. Praesent eget leo efficitur odio bibendum condimentum sit amet vel ex. Nunc maximus quam orci, quis pulvinar nibh eleifend ac. Quisque consequat lacus magna, eu posuere tellus iaculis ac. Sed vitae tortor tincidunt ante sagittis iaculis.

Subsection Two

Nullam mollis tellus lorem, sed congue ipsum euismod a. Donec pulvinar neque sed ligula ornare sodales. Nulla sagittis vel lectus nec laoreet. Nulla volutpat malesuada turpis at ultricies. Ut luctus velit odio, sagittis volutpat erat aliquet vel. Donec ac neque eget neque volutpat mollis. Vestibulum viverra ligula et sapien bibendum, vel vulputate ex euismod. Curabitur nec velit velit. Aliquam vulputate lorem elit, id tempus nisl finibus sit amet. Curabitur ex turpis, consequat at lectus id, imperdiet molestie augue. Curabitur eu eros molestie purus commodo hendrerit. Quisque auctor ipsum nec mauris malesuada, non fringilla nibh viverra. Quisque gravida, metus quis semper pulvinar, dolor nisl suscipit leo, vestibulum volutpat ante justo ultrices diam. Sed id facilisis turpis, et aliquet eros.

Subsubsection Example Duis venenatis eget lectus a aliquet. Integer vulputate ante suscipit felis feugiat rutrum. Aliquam eget dolor eu augue elementum ornare. Nulla fringilla interdum volutpat. Sed tincidunt, neque quis imperdiet hendrerit, turpis sapien ornare justo, ac blandit felis sem quis diam. Proin luctus urna sit amet felis tincidunt, sed congue nunc pellentesque. Ut faucibus a magna faucibus finibus. Etiam id mi euismod, auctor nisi eget, pretium metus. Proin tincidunt interdum mi non interdum. Donec semper luctus dolor at elementum. Aenean eu congue tortor, sed hendrerit magna. Quisque a dolor ante. Mauris semper id urna id gravida. Vestibulum mi tortor, finibus eu felis in, vehicula aliquam mi.

Aliquam arcu turpis, ultrices sed luctus ac, vehicula id metus. Morbi eu feugiat velit, et tempus augue. Proin ac mattis tortor. Donec tincidunt, ante rhoncus luctus semper, arcu lorem lobortis justo, nec convallis ante quam quis lectus. Aenean tincidunt sodales massa, et hendrerit tellus mattis ac. Sed non pretium nibh.

Donec cursus maximus luctus. Vivamus lobortis eros et massa porta porttitor. Nam vitae suscipit mi. Pellentesque ex tellus, iaculis vel libero at, cursus pretium sapien. Curabitur accumsan velit sit amet nulla lobortis, ut pretium ex aliquam. Proin eget volutpat orci. Morbi eu aliquet turpis. Vivamus molestie urna quis tempor tristique. Proin hendrerit sem nec tempor sollicitudin.