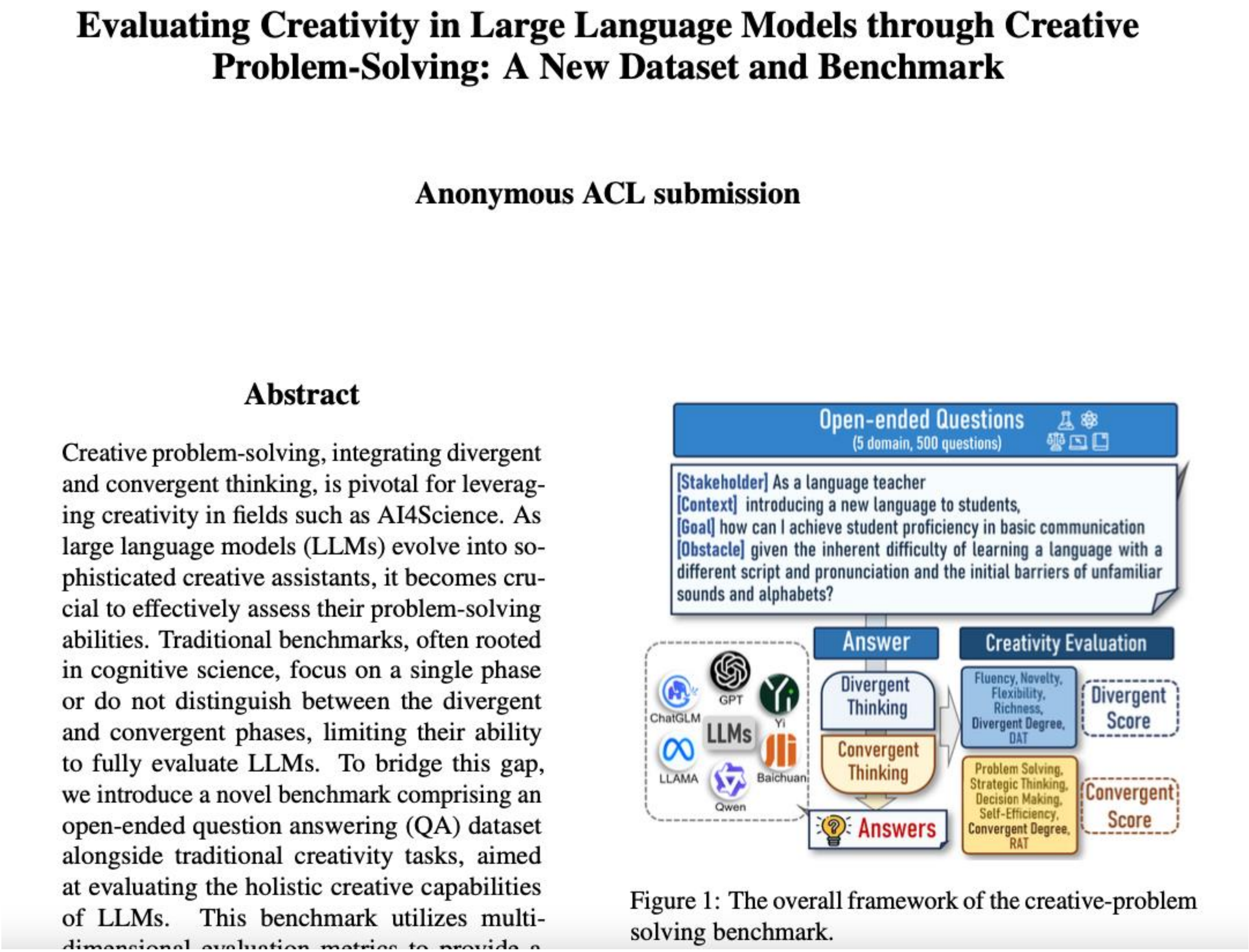
Weekly Sync

5/11/24

Paper Review #2

Sorry guys, my progress update has become another paper review...



Evaluating Creativity in Large Language Models through Creative Problem-Solving: A New Dataset and Benchmark

Anonymous ACL submission

Abstract

Creative problem-solving, integrating divergent and convergent thinking, is pivotal for leveraging creativity in fields such as AI4Science. As large language models (LLMs) evolve into sophisticated creative assistants, it becomes crucial to effectively assess their problem-solving abilities. Traditional benchmarks, often rooted in cognitive science, focus on a single phase or do not distinguish between the divergent and convergent phases, limiting their ability to fully evaluate LLMs. To bridge this gap, we introduce a novel benchmark comprising an open-ended question answering (QA) dataset alongside traditional creativity tasks, aimed at evaluating the holistic creative capabilities of LLMs. This benchmark utilizes multi-dimensional evaluation metrics to provide a

[Stakeholder] As a language teacher [Context] introducing a new language to students, [Goal] how can I achieve student proficiency in basic communication [Obstacle] given the inherent difficulty of learning a language with a different script and pronunciation and the initial barriers of unfamiliar sounds and alphabets?

Open-ended Questions
(5 domain, 500 questions)
ChatGLM
GPT
Yi
LLMs
LLAMA
Qwen
Baichuan
Answer
Divergent Thinking
Convergent Thinking
Answers
Creativity Evaluation
Fluency, Novelty, Flexibility, Richness, Divergent Degree, DAT
Divergent Score
Problem Solving, Strategic Thinking, Decision Making, Self-Efficiency, Convergent Degree, RAT
Convergent Score

Figure 1: The overall framework of the creative-problem solving benchmark.

Why and What This Paper is About?

- **Motivation**
  - Traditional creativity benchmark focuses too much on “divergent thinking” or “convergent thinking”, but a comprehensive measure of both
  - Those does not capture the full spectrum of Creative Problem-Solving (CPS) abilities of LLMs
- **Objectives**
  - Create a new benchmark that measure CPS accurately and effectively
  - Measured through multi-dimensional metrics (will explain later)

Why and What This Paper is About?

- **Methodology**
  - They make a dataset first – this one I will skip because it’s not relevant to what I am doing
  - Under evaluating creativity, there are 3 different tasks:

  1. **Divergent Thinking Assessment (DAT)**
     - Focused on measuring divergent thinking (duh)
     - LLM is prompted to produce a set of unrelated nouns
     - The diversity is measured from the average semantic distance between the generated words
     - The bigger the distance, the more creative the LLM is

Why and What This Paper is About?

- **Methodology**
  - Under evaluating creativity, there are 3 different tasks:

  2. **Remote Associates Test (RAT)**
     - Basically NYT connections
     - LLM is given 3 seemingly unrelated words, and is tasked to find a fourth one that is semantically related to all three
     - e.g. “Power, Friend, Scout” → “Girl”
     - Creativity is calculated from the semantic closeness of the generated word to the ground truth
     - This reflects the convergent thinking, i.e. synthesizing information and make precise connections between concepts

Why and What This Paper is About?

- **Methodology**
  - Under evaluating creativity, there are 3 different tasks:

  3. **Open-ended QA**
     - To assess both convergent and divergent thinking
     - Each question contains multiple possible answers
     - Model is asked to generate multiple distinct solution for each question
     - Divergent thinking is measured by:
       - Fluency, novelty, flexibility, and richness
     - Convergent thinking is measured by:
       - Problem solving, strategic thinking, decision making, and self-efficiency
     - To measure the above, they used LLM-as-a-judge

Why and What This Paper is About?

3. **Open-ended QA**
   - Divergent thinking
     1. **Fluency**: quantifies the volume of ideas
     2. **Novelty**: evaluates the uniqueness
     3. **Flexibility**: assesses the variety across categories
     4. **Richness**: gauges the depth of ideas

   - Convergent thinking
     1. **Problem solving**: assesses how effectively the response address and resolve the issue
     2. **Strategic thinking**: evaluates the response's long-term planning and foresight
     3. **Decision making**: determines the decisiveness and rationale behind the choices made
     4. **Self-efficiency**: judges the confidence and resourcefulness exhibited in the response

Approach & Results

- **Key Findings**
  - Reveals significant differences in CPS abilities across LLMs.
    - GPT-4 is really good in both convergent and divergent tasks
  - There is a positive correlation between model size ad creative performance, with larger models performing better

- **Strengths**
  - A comprehensive framework by integrating divergent and convergent thinking metrics
  - Used innovative metrics such as fluency, novelty and problem solving, adding depth to evaluation
  - ALL PROMPTS ARE OPEN SOURCED (THANK GOODNESS)

Approach & Results

- **Limitations**
  - Focused on convergent and divergent thinking, which still does not capture all dimensions of creativity
  - Relies heavily on empirical metrics and LLM-as-a-judge
  - So there is no deeper theoretical framework (unlike the last paper)

- **Impact**
  - Set a new evaluation standard
  - May lead to the development of a more nuanced and capable LLMs that can better support creative tasks