



Batch Prompting

Expanding Capabilities to Multi-Task and Zero-Shot

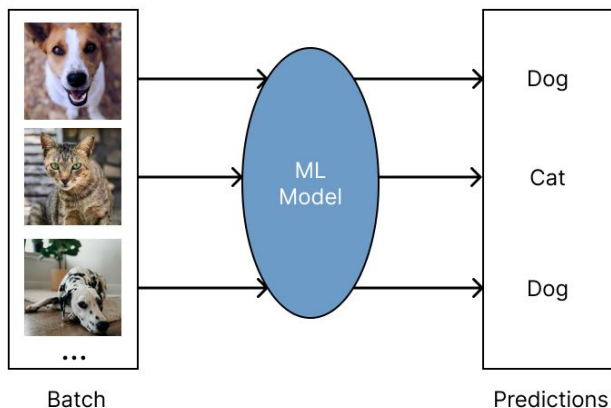
Presenters: Alex Chandler, Rohan Jha

11-30-2023

Motivation

What is Batch Prompting?

Just like batched inference in classic case: multiple-input multiple-output...
but at the language level



Why use Batch Prompting?

Batching samples amortizes the cost of processing a relatively long prefix $\{N_{\text{prompt}}\}$ over $\{b\}$ samples

Tokens per sample:

$$N_{\text{standard}} = N_{\text{prompt}} + N_{\text{response}}$$

$$N_{\text{batched}} = \frac{N_{\text{prompt}}}{b} + N_{\text{response}}$$

[\[ray.io\]](#) [\[Cheng et al. 2023\]](#)

Challenges

- Have LLM maintain consistent output format and answer all questions
 - Requires either a very capable model or lots of prompt engineering
- Interdependence/Confusion
 - Unlike traditional batched inference, sample are not independent in attention/context
 - Prior work [\[Shi et al. \(ICML 2023\)\]](#) shows that “LLMs can be easily distracted by irrelevant context” (such as other questions/answers)

Original Problem

Jessica is six years older than Claire. In two years, Claire will be 20 years old. How old is Jessica now?

Modified Problem

Jessica is six years older than Claire. In two years, Claire will be 20 years old. *Twenty years ago, the age of Claire's father is 3 times of Jessica's age.* How old is Jessica now?

Standard Answer 24

Table 1. An example problem from GSM-IC. An irrelevant sentence (*italic and underlined*) that does not affect the standard answer is added immediately before the question.

Previous Works

January, 2023

Batch Prompting: Efficient Inference with Large Language Model APIs

Zhoujun Cheng

Shanghai Jiao Tong University

blankcheng@sjtu.edu.cn

Jungo Kasai

University of Washington

jkasai@cs.washington.edu

Tao Yu

University of Hong Kong

tyu@cs.hku.hk

Introduced batch technique,
demonstrated reduction in inference
cost with slight performance
degradation

September, 2023

BATCHPROMPT: ACCOMPLISH MORE WITH LESS

Jianzhe Lin, Maurice Diesendruck, Liang Du, Robin Abraham
Microsoft

Explored larger batch sizes ($b=64$),
New iterative voting scheme over
permuted batches

Existing Batch Prompts

Batch Prompting [Cheng et al., 2023]

Chat Prompt for GSM8K

System

The answers should respond to each question the user inputs. One line one answer starting with A[idx], no other line break.

User

Q[1]: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

Q[2]: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

Q[3]: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

Q[4]: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

Assistant

A[1]: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been $21 - 15 = 6$. The answer is 6.

A[2]: There are originally 3 cars. 2 more cars arrive. $3 + 2 = 5$. The answer is 5.

A[3]: Originally, Leah had 32 chocolates. Her sister had 42. So in total they had $32 + 42 = 74$. After eating 35, they had $74 - 35 = 39$. The answer is 39.

A[4]: Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny $20 - 12 = 8$. The answer is 8.

User

Q[1]: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

Q[2]: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

Q[3]: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?

Q[4]: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?

Assistant

BatchPrompt [Lin et al., 2023]

Prompts for GSM8K:

You will be given [BATCH-SIZE] math problems.

These problems take between 2 and 8 steps to solve, and solutions primarily involve performing a sequence of elementary calculations using basic arithmetic operations to reach the final answer.

The below is the format of sentence pairs which will be given:

=====

Input 0: xxxxx

Input 1: xxxxx

.....

=====

Below are the calculation results you need to generate. [Conf-Description]

===== Result for Input 0: [intermediate reasoning steps], The answer is xxxx. [Place-holder-Conf]

Result for Input 1: [intermediate reasoning steps], The answer is xxxx. [Place-holder-Conf]

.....

=====

Please make sure to write your a series of intermediate reasoning steps. Please make sure the final sentence is "The answer is xxx.", and the answer should be a number.

Please make sure to generate [BATCH-SIZE] labels each time.

Prompt then includes few-shot examples (omitted to fit on screen)

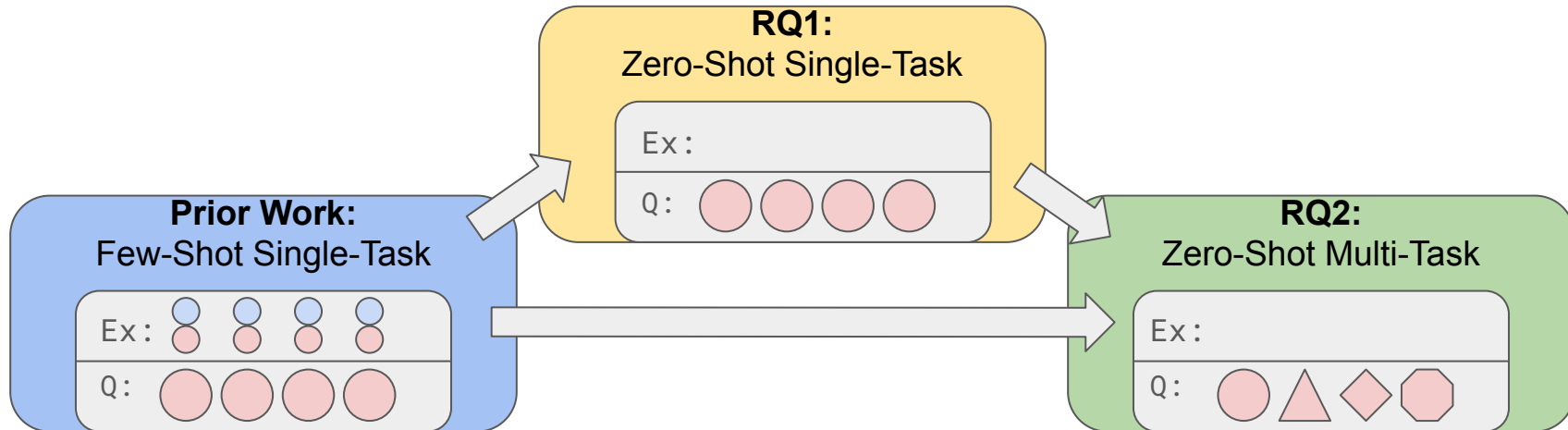
Our Research Questions

1. Example-Dependent Efficacy

- Can Batch Prompting work in a Zero-Shot setting?

2. Multi-Task Batch Prompting

- Can Batch Prompting handle a batch of unrelated queries effectively?
- How does performance of individual queries vary in mixed (diverse) vs. related query batches?



Our Batch Prompts

- Development of Our Batched Prompts:
 - Prompts Created and Tested in OpenAI Playground with GPT-4 as model.
- Key Components of Batched Prompts:
 - Unambiguous Task Description
 - Input & Output Format Specifications
 - Completeness Reminder

0.1 Single-Task BatchPrompt - GSM8K Zero-Shot

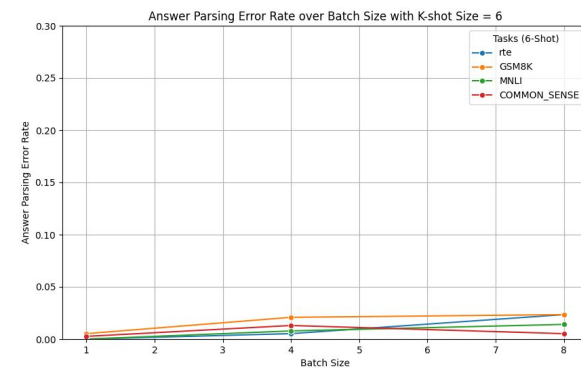
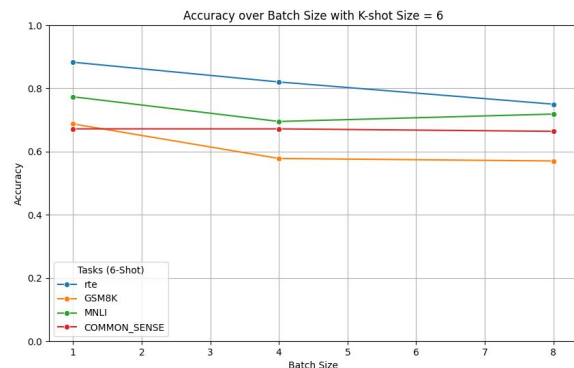
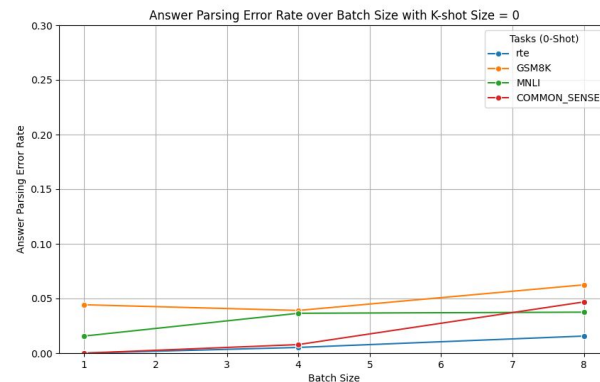
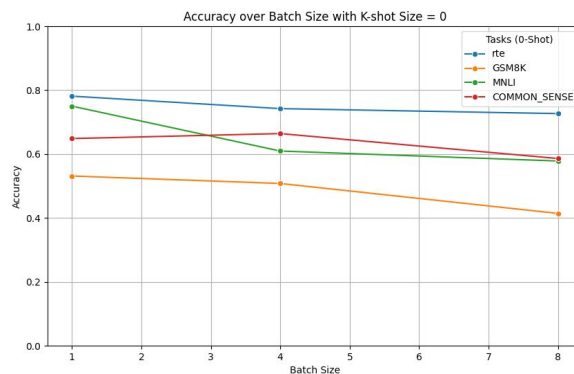
Batch Prompt Example
Objective: Your task is to solve a set of math questions in a batch. The total number of questions in the batch is defined as {batch_size}.
Complexity: Each question in the batch will require you to perform between 2 and 8 steps to arrive at the final answer.
Method: Use basic arithmetic operations to perform a sequence of calculations for solving these questions.
Instructions: 1. Intermediate Reasoning: Include all the intermediate steps you took to arrive at each answer. This ensures that the answer is both well-reasoned and reliable. 2. Batch Size: The number of answers you provide must exactly match the specified {batch_size}.
Input Format: Q[0]: {Question_0.Text} Q[1]: {Question_1.Text} ... Q[{batch_size - 1}]: {Question_{batch_size - 1}.Text}
Output Format: A[index]: {Intermediate Reasoning}; The answer is {Answer_Integer}
• index: This is the index of the question you are answering. It must be prefixed with 'A' and enclosed in square brackets. • {Intermediate Reasoning}: This is where you provide all the intermediate steps that led you to the final answer. • {Answer_Integer}: This is the final integer answer to the question.
The phrase 'The answer is' must directly precede the integer answer and come after the intermediate reasoning, separated by a semicolon. Please adhere strictly to these guidelines to ensure the output is in the desired format.

0.1 Multi-Task BatchPrompt - Various Domains

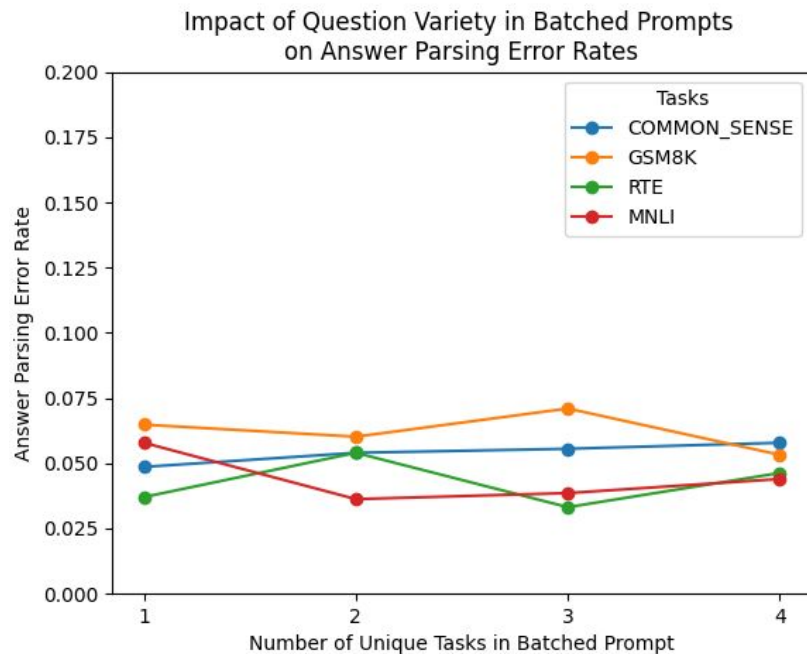
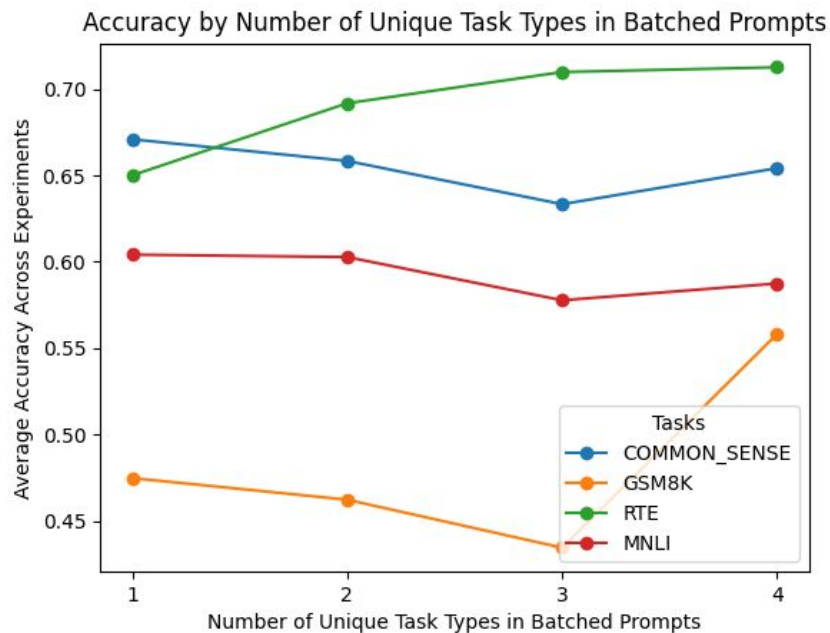
Batch Prompt Example
Objective: Your task is to solve a variety of questions across multiple domains in a single batch operation. The total number of questions in the batch to answer is defined as batch_size = 4.
Task Descriptions: COMMON_SENSE: Your task is to solve a set of multiple-choice questions from the CommonsenseQA dataset in a batch. <Rest of task description excluded to fit on slide> GSM8K: Your task is to solve a set of math questions in a batch. <Rest of task description excluded to fit on slide> RTE: Your task is to solve a set of recognizing textual entailment (RTE) questions in a batch. <Rest of task description excluded to fit on slide> MNLI: Your task is to solve a set of MultiNLI (MNLI) questions in a batch. <Rest of task description excluded to fit on slide>
Instructions: 1. Intermediate Reasoning: Include all the steps you took to evaluate the relationship between the Premise and Hypothesis. 2. Batch Size: You must provide an answer for each question in the batch, ensuring that the number of answers you provide exactly matches the specified 4. 3. Handling Ambiguities: Answer every question even if you are unsure about the answer.
Unified Input Format: The batch will contain questions each tagged with a unique index and a task code, formatted as follows: Q[index task_code]: {{Question.Text}}
Output Format: You must adhere to the following format rigorously for each answer: A[index]: {{Intermediate.Reasoning}}; The answer is {{Answer.Integer}}

Evaluation - Effectiveness of Single Task Batch Prompts for Zero-Shot

Results for GPT
3.5 Turbo 16K:



Evaluation - Effectiveness of Multi Task Batch Prompts for Zero-Shot



Batch Size: 4 questions, Model: GPT-3.5-turbo-16k, Sample Size: 240 questions per dataset, Temperature Setting: 0

Summary

- Demonstrated Feasibility of Zero-Shot Batched Prompting:
 - Successfully implemented batched prompting without exemplars, maintaining low error rates.
- Demonstrated Feasibility of Multi-Task Batched Prompts
 - Effective Integration of Varied Tasks: Demonstrated that batch prompts can integrate multiple distinct tasks while preserving task-specific accuracy.
 - Zero-Shot Capability: Achieved low error rates in parsing within a Zero-Shot setting
- Release of Batched Prompts for Single-Task and Multi-Task Setting

Limitations and Future Works

- Limitations

- Increased Answer Parsing Error Rate in Zero-Shot setting
- **Lots** of prompt engineering required
- Limited domain in which batch prompting achieves greatest efficiency benefit
 - Ideal case: long input, short output

$$N_{\text{batched}} = \frac{N_{\text{prompt}}}{b} + N_{\text{response}}$$

- Extensions

- How to balance diverse/relevant exemplars when there are multiple questions in a batch?
- How does model-size impact batch prompting performance?

End of Presentation

Thank you for your attention! Any questions?

Appendix

- The following slides are to help answer questions as part of QA and provide more detail about our experiments

Extra Detail on Multi-Task Batched Experiments

