Baby's CoThought: Leveraging Large Language Models for Enhanced Reasoning in Compact Models

Zheyu Zhang, Han Yang, Bolei Ma, David Rügamer, Ercong Nie

Abstract

Large Language Models (LLMs) demonstrate remarkable performance on a variety of natural language understanding (NLU) tasks, primarily due to their in-context learning ability. This ability could be applied to building baby-like models, i.e. models at small scales, improving training efficiency. In this paper, we propose a **CoThought** pipeline, which efficiently trains smaller "baby" language models (BabyLMs) by leveraging the **C**hain **o**f Thought prompting of LLMs. Our pipeline restructures a dataset of less than 100M in size using *GPT-3.5-turbo*, transforming it into task-oriented, human-readable texts that are comparable to the school texts for language learners. The BabyLM is then pretrained on this restructured dataset in a *RoBERTa* fashion. In evaluations across 4 benchmarks, our BabyLM outperforms the vanilla *RoBERTa* in 10 linguistic, NLU, and question-answering tasks by more than 3 points, showing a superior ability to extract contextual information. These results suggest that compact LMs pretrained on small, LLM-restructured data can better understand tasks and achieve improved performance.

CoThought Pipeline

Creative NLU-Example Generation:

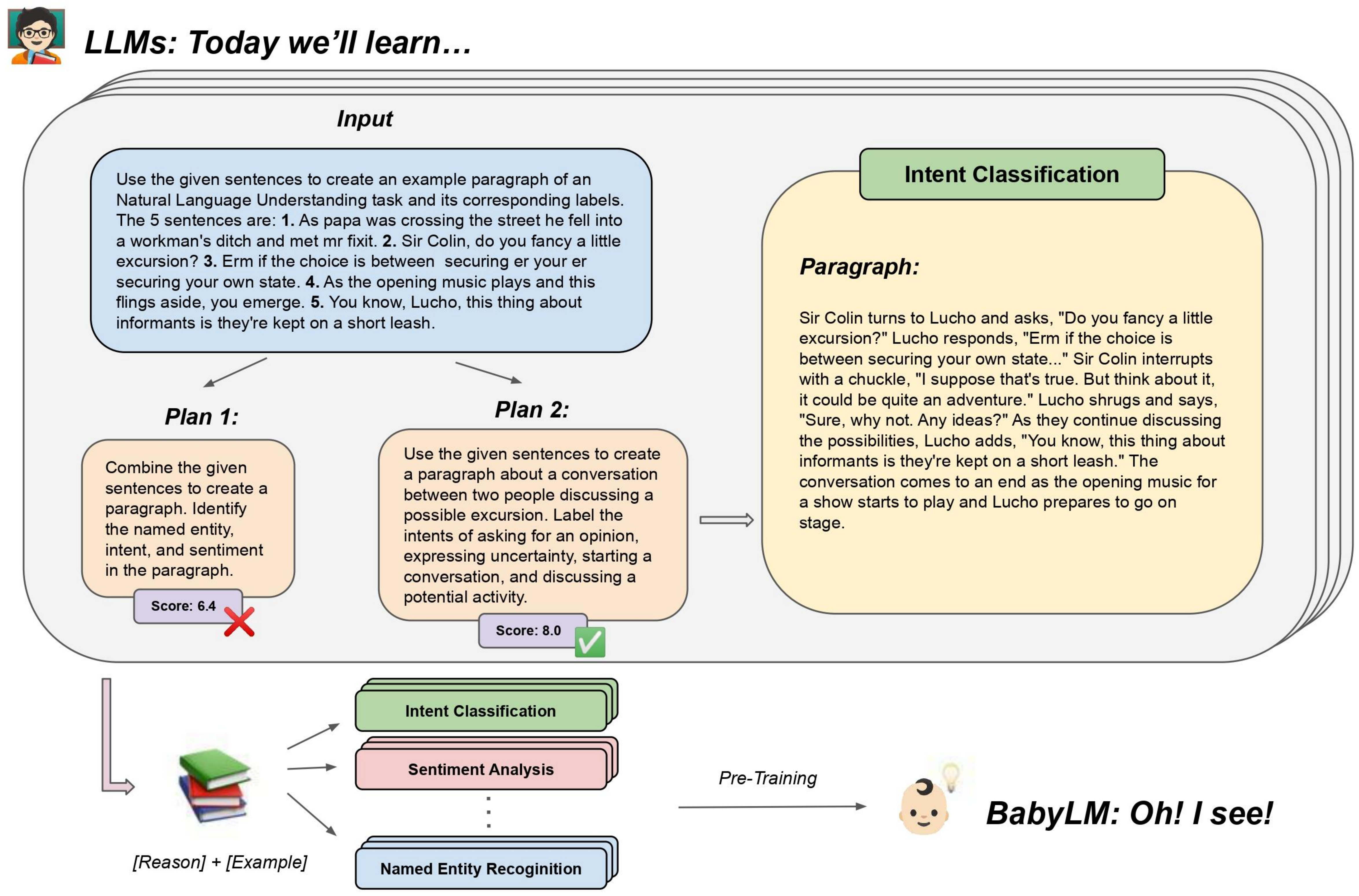
- clean the dataset which contains single sentences
- randomly sample five unique sentences from the dataset
- provide a specific CoT prompt P to let LLM create distinct plans for NLU task generation, each of which is scored based on a coherence scoring prompt P_S of LLM
- based on the best scored plan, use the provided sentences to compose an example paragraph that illustrates a possible intrinsic NLU task and generate labels for the task

Training Data Construction:

- save all the generated paragraphs with labels of NLU tasks
- In order to construct a high-quality dataset for pretraining, set up a threshold for the coherence scoring prompt P_S, and filter out all the texts with a score lower than this threshold

Pretraining:

- pretrain a baby LM based in a vanilla LM (*RoBERTa base*) pretraining on the new constructed dataset



Use the given sentences to create an example paragraph of an Natural Language Understanding task and its corresponding labels. The 5 sentences are: **1.** As papa was crossing the street he fell into a workman's ditch and met mr fixit. **2.** Sir Colin, do you fancy a little excursion? **3.** Erm if the choice is between securing er your er securing your own state. **4.** As the opening music plays and this flings aside, you emerge. **5.** You know, Lucho, this thing about informants is they're kept on a short leash.

Combine the given sentences to create a paragraph. Identify the named entity, intent, and sentiment in the paragraph.

Use the given sentences to create a paragraph about a conversation between two people discussing a possible excursion. Label the intents of asking for an opinion, expressing uncertainty, starting a conversation, and discussing a potential activity.

Sir Colin turns to Lucho and asks, "Do you fancy a little excursion?" Lucho responds, "Erm if the choice is between securing your own state..." Sir Colin interrupts with a chuckle, "I suppose that's true. But think about it, it could be quite an adventure." Lucho shrugs and says, "Sure, why not. Any ideas?" As they continue discussing the possibilities, Lucho adds, "You know, this thing about informants is they're kept on a short leash." The conversation comes to an end as the opening music for a show starts to play and Lucho prepares to go on stage.

Experimental Setups

Dataset:

- BabyLM_100M dataset (Warstadt et al., 2023) which contains simple subsets

Data Generation via CoT Prompting:

- We leveraged the API of *GPT-3.5-turbo* from OpenAI and provided CoT prompt with the format:
  - *Use the given sentences to create an example paragraph of an NLU task and its corresponding labels. The 5 sentences are: input.*
  - *Make a plan then write and determine. Your output should be of the following format:*
  - *Plan:*
  - *Your plan here.*
  - *Paragraph:*
  - *Your paragraph here.*
  - *Task:*
  - *[Only the task name here, without additional information.]*
  - *Labels:*
  - *[Only the labels here, without additional information.]*
- To evaluate the task plans, we prompt the GPT again with the score prompt with the following format, and save the plans which has a score bigger than 7:
  - *Analyze the following paragraph, then at the last line conclude "Thus the coherency score is s", where s is an integer from 1 to 10*

Pretraining:

- We trained a RoBERTa model on the generated dataset using *RobertaForMaskedLM* provided by the *huggingface* library

A case of generated data:

- Paragraph:
  - We have a few topics to cover in this paragraph. Firstly, a possible I.D. has been found in one of Gina's snapshots. Secondly, there is a new technology in development called autostereoscopic 3D that will allow people to watch 3D movies without glasses. This is great news for those who find wearing 3D glasses uncomfortable and causes eye strain. Unfortunately, the narrator regrets not asking Jean for the details about something. Lastly, the police are seen moving down the main street of Atenco, and we are tracking their movements.

- Plan:

✓ 1. Introduce the topic of the paragraph
2. Mention the possible I.D. from Gina's snapshots
3. Talk about the new technology called autostereoscopic 3D
4. Mention the difficulty of wearing 3D glasses
5. Mention the regret of not asking Jean for details
6. Talk about the police and their movement down the main street of Atenco

- Task:
  - Text Classification

- Labels:

1. I.D. Mentioned
2. Technology Mentioned
3. Regret Expressed
4. Police Mentioned

Performance on selected benchmarks:

Tasks	Our BabyLM	Vanilla LM	Difference
BLiMP			
Filler Gap	**78.52**	68.00	**10.52**
Subject Verb Agreement	**85.17**	76.20	**8.97**
Argument Structure	**78.06**	71.30	**6.76**
Determiner Noun Agreement	**97.75**	93.10	**4.65**
Anaphor Agreement	**93.61**	89.50	**4.11**
Ellipsis	77.02	83.80	-6.78
Island Effects	45.85	54.50	-8.65
BLiMP Supplement			
Subject Aux Inversion	**77.73**	45.60	**32.13**
QA Congruence Easy	**62.50**	34.40	**28.1**
Turn Taking	**62.50**	46.80	**15.7**
GLUE			
BoolQ	**65.84**	59.90	**5.94**
MNLI	**73.73**	68.70	**5.03**
MNLI-mm	74.76	78.00	-3.24
QNLI	76.86	82.30	-5.44
RTE	45.45	51.50	-6.05
AVG. (overall)	**73.95**	71.75	**2.20**

Table content: selected results of our BabyLM and the vanilla LM *RoBERTa*, where the performance of BabyLM has been improved by at least 3 points (in bold), or reduced (-) over 3. The metric in this table is all accuracy score.

Takeaways:

→ LLM is able to reformulate raw data into reduced simple texts of NLU tasks by its CoT reasoning prompting

→ The BabyLM trained on the LLM-restructured and -inferred data achieves higher performance in many linguistic tasks compared to vanilla LMs, even in the case of small training data volume.

Next Steps:

→ Use different LLMs to generate the pretraining data and compare the difference

→ Try a broader range of architectures, including causal language models and various transformer variants, for pretraining

✉ Presenting authors:
bolei.ma@lmu.de
nie@cis.lmu.de

Paper @ arXiv Code @ Model @ Data @

BabyLM Challenge
CoNLL 2023
EMNLP 2023
gesis Leibniz-Institut für Sozialwissenschaften
MCML Munich Center for Machine Learning
CIS
LMU
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN