семинар на кафедре ММП ВМК МГУ

материал на основе выступления на семинаре Dasha.AI

Разметка с помощью GPT-3 и «Active labeling»

Александр Дьяконов

13 октября 2021 года

«Хотите уменьшить стоимость разметки? GPT-3 может помочь»

Want To Reduce Labeling Cost? GPT-3 Can Help

Shuohang Wang Yang Liu Yichong Xu Chenguang Zhu Michael Zeng

Microsoft Cognitive Services Research Group

{shuowa,yaliu10,yicxu,chezhu,nzeng}@microsoft.com

Abstract

Data annotation is a time-consuming and labor-intensive process for many NLP tasks. Although there exist various methods to produce pseudo data labels, they are often task-specific and require a decent amount of labeled data to start with. Recently, the immense language model GPT-3 with 175 billion parameters has achieved tremendous improvement across many few-shot learning tasks. In this paper, we explore ways to leverage GPT-3 as a low-cost data labeler to train other models. We find that, to make the downstream model achieve the same performance on a variety of NLU and NLG tasks, it costs 50% to 96% less to use labels from GPT-3 than using labels from humans. Furthermore, we propose a novel framework of combining pseudo labels from GPT-3 with human labels, which leads to even better performance with limited labeling budget. These results present a cost-effective data labeling methodology that is generalizable to many practical applications.

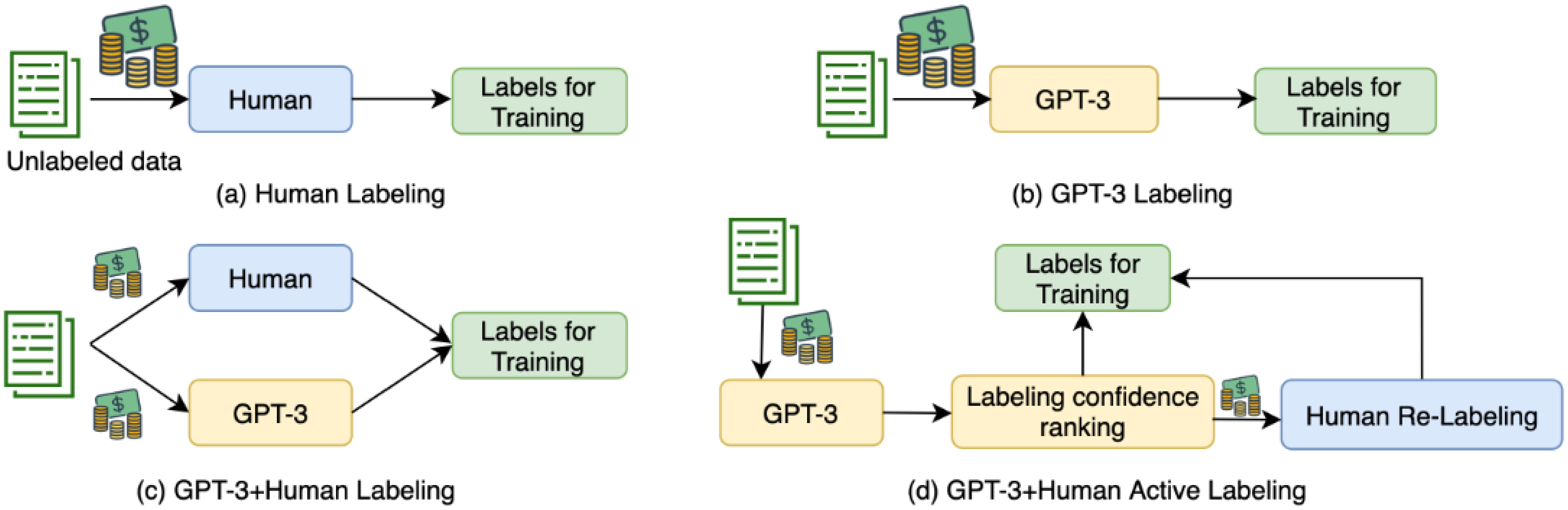
of processed tokens[1]. Thus, an interesting problem arises: instead of directly deploying GPT-3 for downstream tasks, how can we leverage GPT-3 to achieve a more cost-effective and efficient training of other models?

In this paper, we employ GPT-3 to label unannotated data to train smaller models which are deployed for inference. Although the data labeled by GPT-3 is usually more noisy than human-labeled data, the process is much cheaper, faster and generalizable to multiple tasks. For example, for the Stanford Sentiment Treebank (SST-2) task (Socher et al., 2013), it takes as low as 0.002 dollars on average to use the GPT-3 API to annotate one label. However, it costs 0.11 dollars to label an instance on crowd-sourcing platforms. Plus, the GPT-3 API can label data non-stoppingly at a much faster speed than human labelers.

In our extensive empirical analysis, we find that to make in-house models (e.g. PEGASUS (Zhang et al., 2020), RoBERTa (Liu et al., 2019)) to achieve the same performance on various NLU

https://arxiv.org/pdf/2108.13487.pdf

Схемы разметки



Разметка с GPT-3 дешевле: 0.002$ – 0.11$
и быстрее

В (d) человек переразмечает то, в чём низкая уверенность – «Active labeling»
(разные проценты 0%, 25%, 50%, 75%, 100% пробовали)

Разметка с помощью OpenAI

Pricing

Simple and flexible. Only pay for what you use.

JOIN THE WAITLIST ›

Per-model prices

The API offers multiple models with different capabilities and price points. Davinci is the most powerful model, while Ada is the fastest.

Prices are per 1,000 tokens. You can think of tokens as pieces of words, where 1,000 tokens is about 750 words. This paragraph is 35 tokens.

Learn more

MODEL		PRICE PER 1K TOKENS
Davinci	Most powerful	$0.0600
Curie		$0.0060
Babbage		$0.0012
Ada	Fastest	$0.0008

Используют GPT-3 API from OpenAI (top-k predicted tokens at each output position)

https://beta.openai.com/pricing

Разметка с помощью GPT

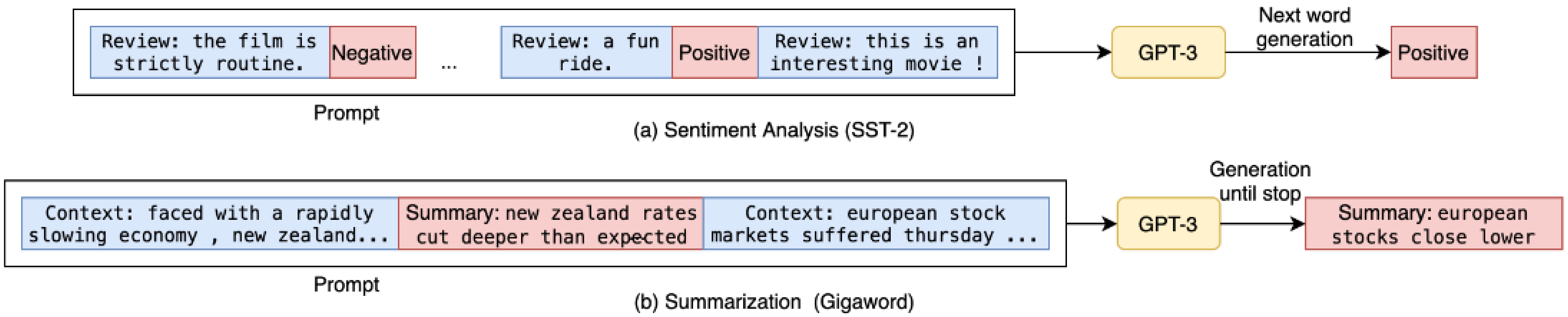


Figure 1: Two examples of constructing GPT-3 input. The input prompt of GPT-3 consists of n labeled data (n-shot learning) and the task input for which GPT-3 generates the label. The same n labeled data is used for every input.

несколько примеров разметки + что надо разметить → метка

Использование разметки

на такой разметке обучили две модели

PEGASUS (Zhang et al., 2020) for NLG
RoBERTalarge (Liu et al., 2019) for NLU
инициализация из оригинальных работ
версии «Large»

такие модели могут превосходить GPT3!

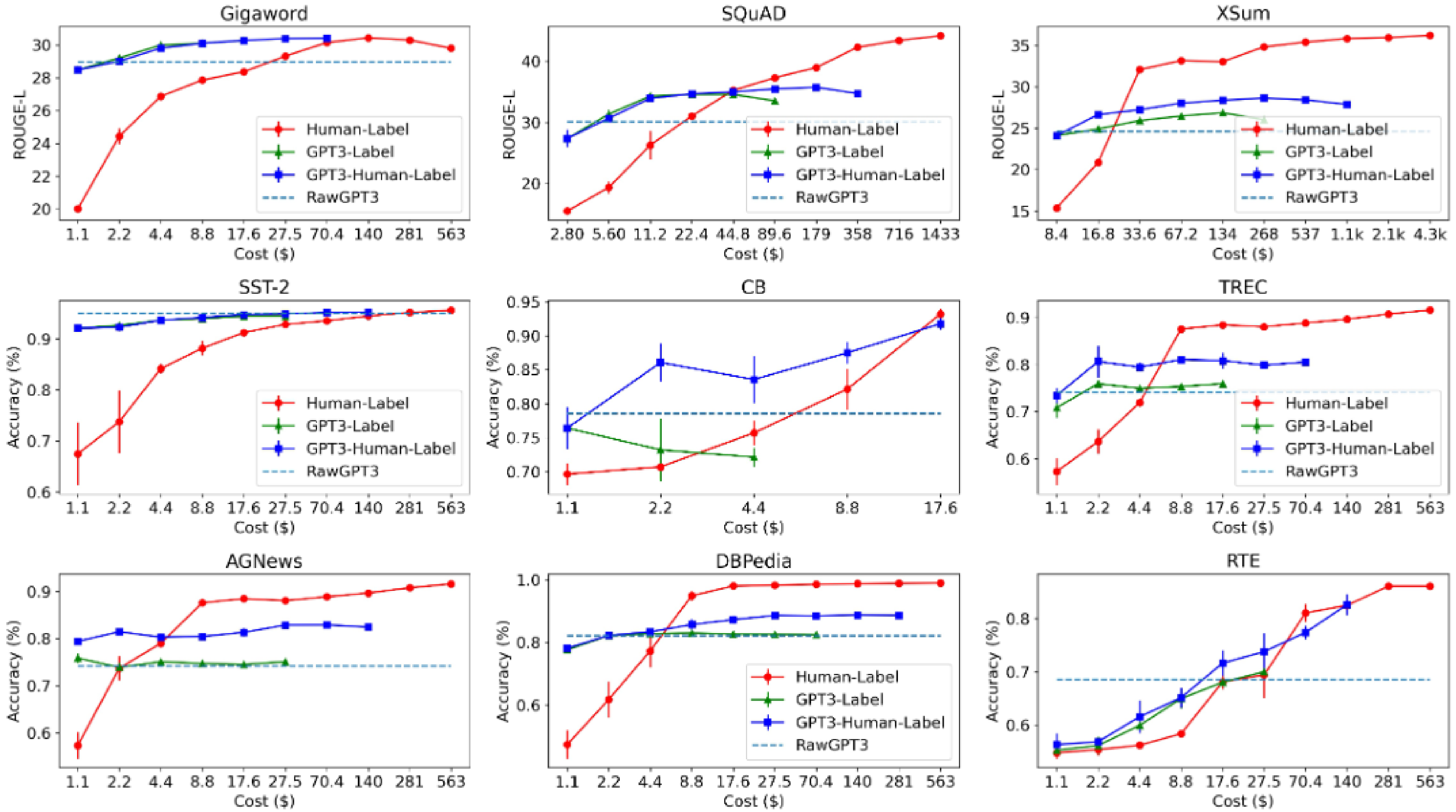


Figure 3: Performance v.s. labeling cost of various labeling strategies on 9 NLG and NLU datasets. X-axis is the cost in dollar estimated by OpenAI pricing policy and crowd-sourced annotation. Each point is the average result of 3 runs of PEGASUS (NLG) or RoBERTa$_{large}$ (NLU) using 3 sets of generated labels, with the standard deviation shown. The performance of using GPT-3 as the inference model is shown as a dashed line, which is the maximum ROUGE-L/accuracy over different shot settings. Note that the cost of GPT3-Label and GPT3-Human-Label cannot further increase when all training data (up to 5,120 instances) has been labeled.

Эксперименты

Но тут «human labeling» – симуляция (взяли метки из датасетов)
Каждый раз фиксированое (5.1K) число объектов

В условиях низкого бюджета автоматическая разметка лучше

На следующем слайде пунктиром – применение чистой GPT для указанных задач

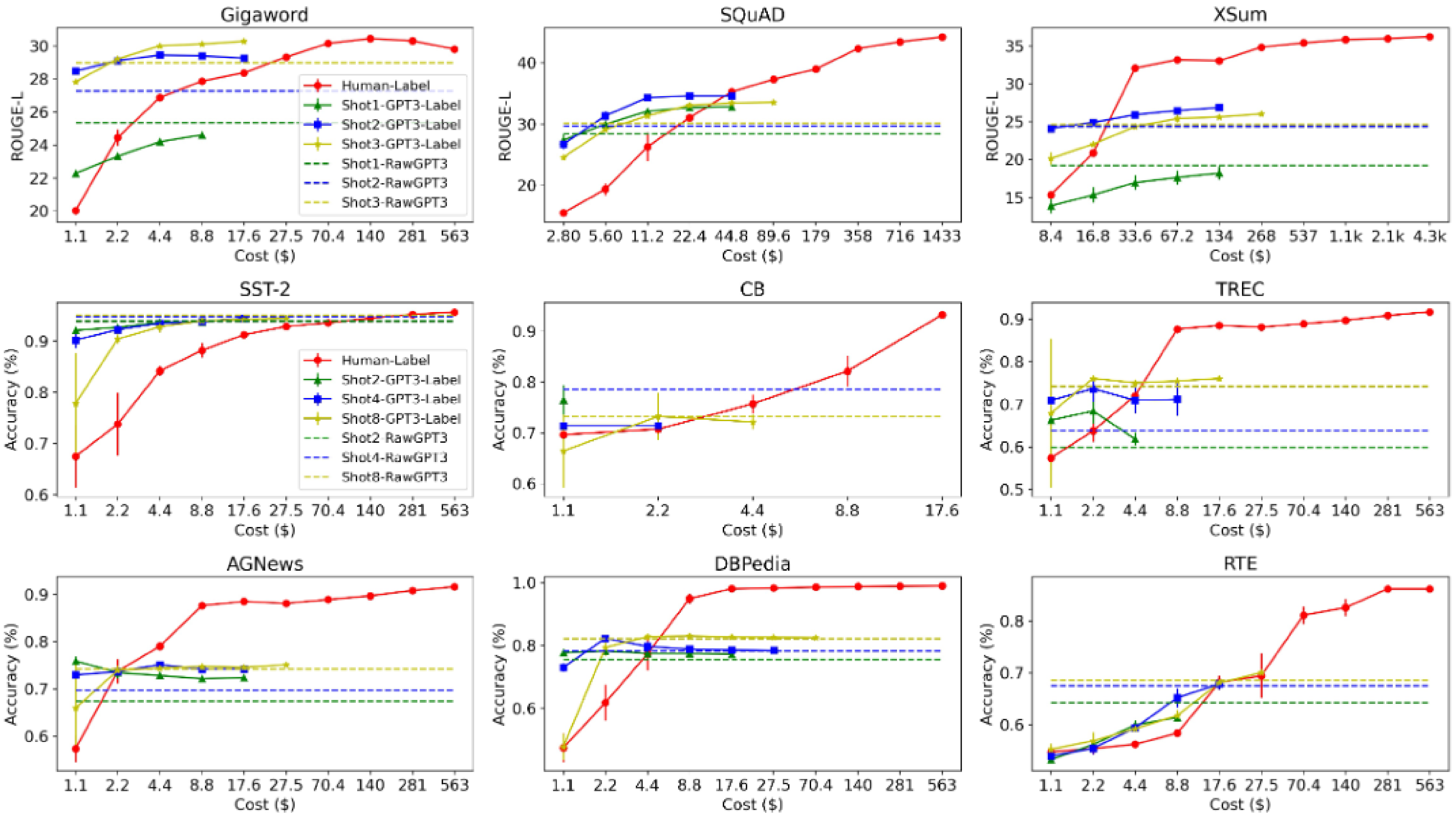


Figure 4: GPT-3 labeling performance. We feed un-labeled data to GPT-3 with different shot settings and fine-tune Transformer models on the corresponding labeled data. The dot lines are the raw GPT-3 performance with various shots. Lines in the same color use the same number of shots in GPT-3. The cost of GPT3-Label cannot further increase when all training data (up to 5,120 instances) has been labeled.

Эксперименты с активной разметкой

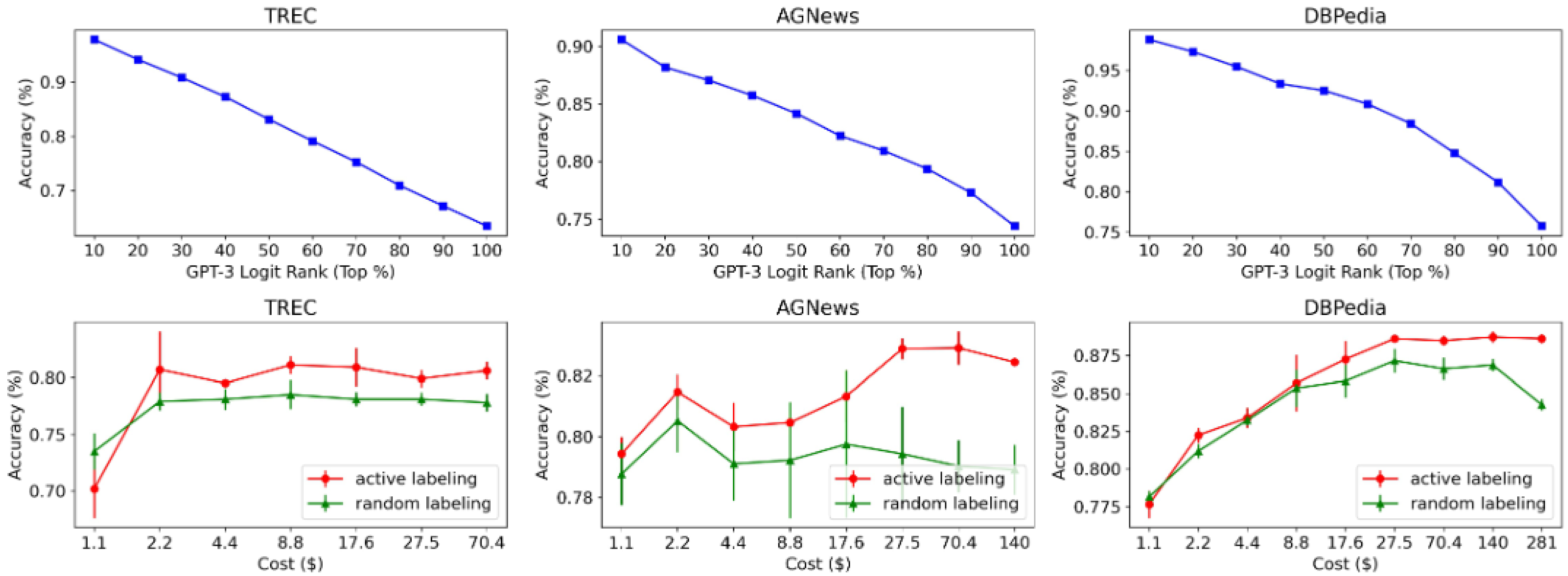


Figure 5: Active labeling. The first row shows that logit values from GPT-3 can be treated as confidence scores, and high-confidence labels are much more accurate than low-confidence ones. The second row compares the performance of active labeling and random labeling in GPT3-Human strategy on three different NLU datasets.