UNICAMP

Article Presentation

IA368DD_2023S1: Deep Learning aplicado a Sistemas de Buscas

Student: Marcus Vinícius Borela de Castro

Visconde e WebGPT

- WebGPT: Browser-assisted question-answering with human feedback (https://arxiv.org/pdf/2112.09332.pdf)

- Visconde: Multi-document QA with GPT-3 and Neural Reranking (https://arxiv.org/pdf/2212.09656.pdf)

A rising challenge in NLP is long-form question-answering (LFQA), in which a paragraph length answer is generated in response to an open-ended question whose supporting evidence is spread over multiple (potentially long) documents.

Main concepts

Visconde

The name is a homage to Visconde de Sabugosa a fictional character that is a corn cob doll whose wisdom comes from reading books.

Motivations:
The few-shot capability of LLMs may reduce the costs for solving QA tasks, as it allows one to implement QA systems for different domains without needing a specific annotated dataset
Recent studies showed that adding a chain-of-thought (CoT) reasoning step before answering significantly improves LLMs' zero or few-shot effectiveness on diverse QA benchmarks.

Visconde Flow

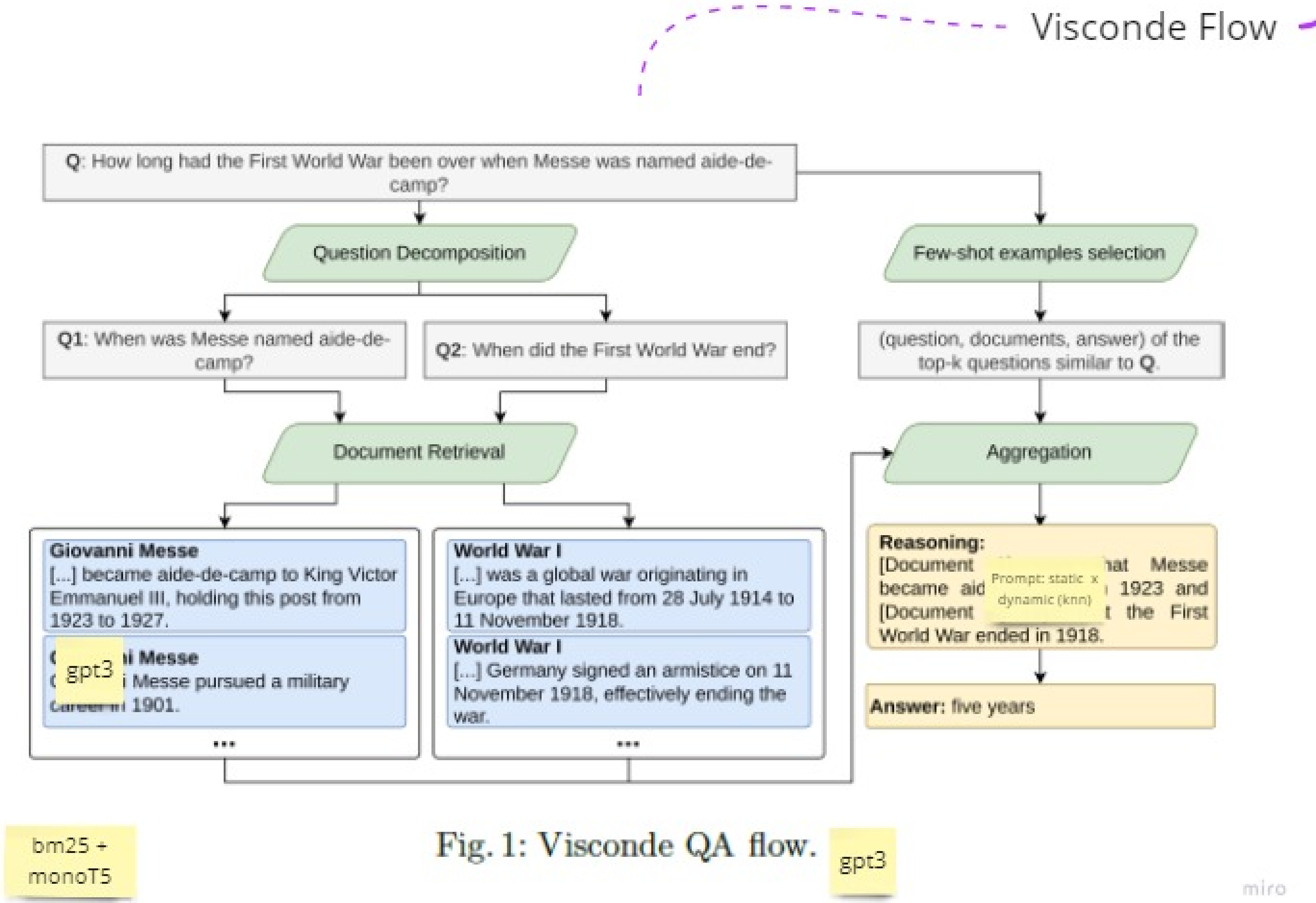


Fig. 1: Visconde QA flow.

Aggregagtion tests:
(**Gold Ctx**) skipping the retrieval step with and without CoT
(**Gold linked pages**) using the reasoning step with the links and ground truth contexts
(**All linked pages**) using reasoning step over the intersection of retrieved documents and the documents cited by the main context
(**Retrieaval**) reasoning over the documents retrieved from the entire Wikipedia subset provided by the dataset

WebGPT

By setting up the task so that it can be performed by humans, we are able to train models on the task using imitation learning, and then optimize answer quality with humanfeedback.To make human evaluation of factual accuracy easier, modelsmust collect references while browsing in support of their answers.

Guiddance from human is central to the approach: (**human feedback**):
demonstrations (6k) of humans using our web-browsing environment to answer questions,
comparisons (21k) between two model-generated answers to the same question
Answers are judged for their factual accuracy, coherence, and overall usefulness

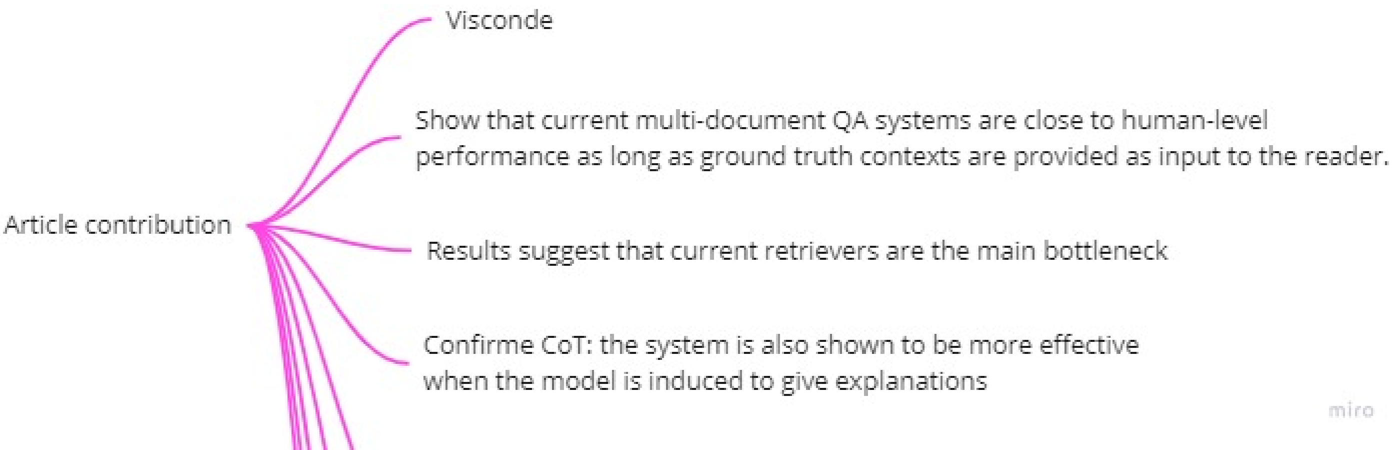
Use this data in four main ways:

behavior cloning BC (i.e., supervised fine-tuning) using the demonstrations

reward modeling RM using the comparisons

reinforcement learning RL against the reward model

rejection sampling RS against the reward model


Article contribution
Visconde
Show that current multi-document QA systems are close to human-level performance as long as ground truth contexts are provided as input to the reader.
Results suggest that current retrievers are the main bottleneck
Confirme CoT: the system is also shown to be more effective when the model is induced to give explanations
miro

WebGPT

use human feedback to directly optimize answer quality,
allowing us to achieve performance competitive with humans.

focus on the higher-level task of using a search engine to answer questions, something that humans can do well, and that a language model can mimic: create a text-based web-browsing environment that a fine-tuned language model can interact with. This allows to improve both retrieval and synthesis in an end-to-end fashion using general methods such as imitation learning and reinforcement learning.

Transparency. It is much easier to understand how WebGPT composes answers than it is for GPT-3, since the entire browsing process can be inspected. It is also straightforward for end-users to follow up on sources to better judge factual accuracy for themselves.

The traditional methods formulate retrieval as a differentiable process. Fully differentiable retrieval has the advantage of fast optimization... But two disadvantages (overcome by webGPT) are that it cannot deal with non-differential processes like using a search engine, and it is less interpretable.

Interesting/unexpected results

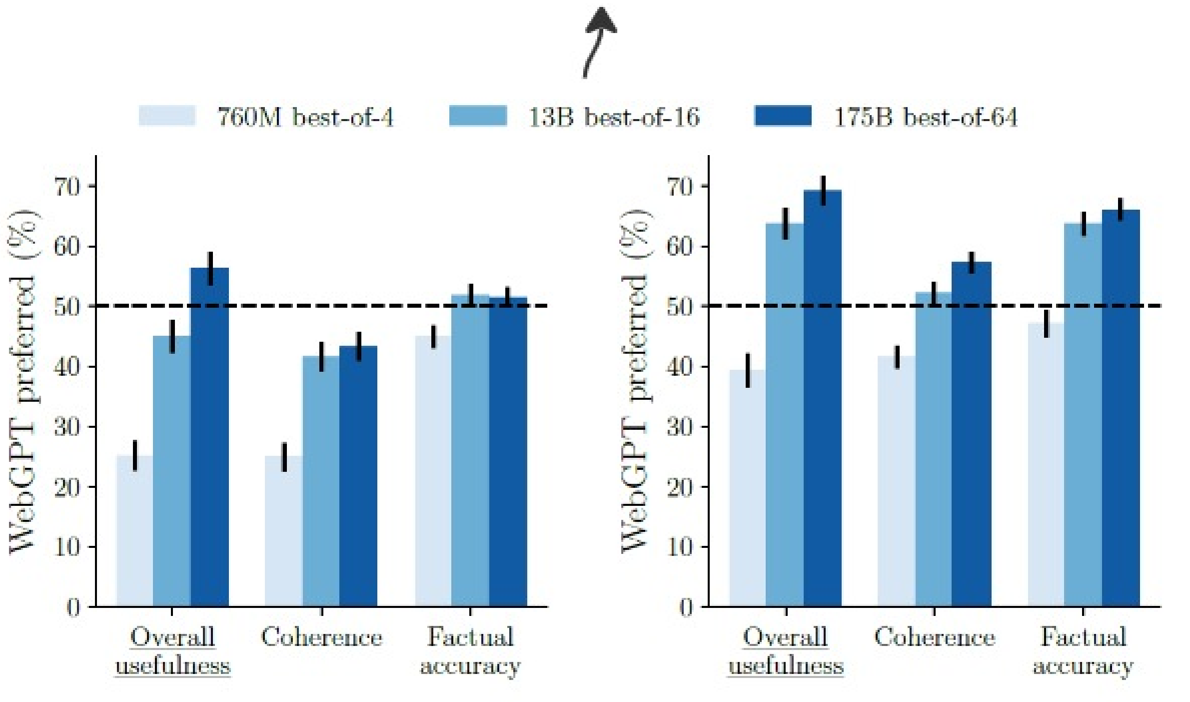
Visconde rivals state-of-the-art supervised models in three datasets: IIRC, Qasper, and StrategyQ

IIRC		
Model	F1	EM
Human	88.4	85.7
Finetuned		
Ferguson et al. [7]	31.1	27.7
Ferguson et al. [7] Linked pages	32.5	29.0
Ferguson et al. [7] Gold Ctx	70.3	65.6
PReasM (pretrain + finetuning) [30] Gold Ctx	-	73.3
PReasM (pretrain + finetuning) [30]	-	47.2
Sup_{A+QA} (supervised) [8]	51.6	-
Few-shot		
Visconde (4-shot dynamic prompt) Gold Ctx and CoT	84.2	74.7
Visconde (4-shot dynamic prompt) Gold Ctx	80.3	70.0
Visconde (4-shot static prompt) Gold Ctx	74.3	62.7
Visconde (4-shot dynamic prompt) Linked pages	48.2	40.7
Visconde (4-shot dynamic prompt) CoT	47.9	40.0

WebGPT

the percentage of truthful and informative answers increases with model size for WebGPT, unlike GPT-3 (TruthfulQA data)

tHE est model, the 175B best-of-64 model, produces answers that are preferred to those written by our human demonstrators 56% of the time (ELI5 DATA)



(a) WebGPT vs. human demonstrations.

(b) WebGPT vs. ELI5 reference answers.

Basic doubts that may arise

Why to remove "the final unembedding layer" in RM method?

The purpose of removing the last unembedding layer is to modify the model's architecture for a specific purpose. By removing the final unembedding layer, the model's output **becomes a latent representation** that can be used as input for further training or processing, such as predicting rewards in the reward modeling step (RM in WebGPT).   (by ChatGPT)

What is reinforcement learning and the the Proximal Policy Optimization (PPO) algorithm used by Reinforcement Learning (RL) method in webGPT?

By training the model using RL, it learns to make better decisions and optimize its behavior based on the rewards obtained from the reward model, ultimately enhancing its performance in the targeted environment.

RL is a branch of machine learning where an agent learns to make sequential decisions interacting with an environment, observing its state, taking actions, and receiving rewards based on its actions.
The agent's objective is to learn a policy, which is a mapping from states to actions, that maximizes the expected cumulative reward over time.

PPO is designed to optimize and update the agent's policy iteratively without explicitly modeling the environment dynamics. It utilizes a "proximal" objective function that provides stability during the policy updates.
(by ChatGPT)

Advanced topic to discuss

There are a number of ways in whichWebGPT tends to perpetuate and reinforce existing assumptions and biases: inherits the biases of the base model; reinforce and entrench existing beliefs and norms as it synthesizes info from existing sources; accepts the implicit assumptions made by questions, and more generally seems to be influenced by the stance taken by questions

These problems could be mitigated with improvements both to WebGPT's base model and to WebGPT's training objective ***How???***

Risks of live web access

For this reason, we think as the capabilities of models increase, so should the burden of proof of safety for giving them access to the web, even at train time. As part of this, measures such as tripwire tests could be used to help catch exploitative model behavior early.

What are possible tripwire tests?

Extra: A little about [CoT (Chain of Thought)]

A prompting only approach is important because it does not require a large training dataset and because a single model checkpoint can perform many tasks without loss of generality

the article show results on arithmetic reasoning, commonsense reasoning, and symbolic reasoning.

Example

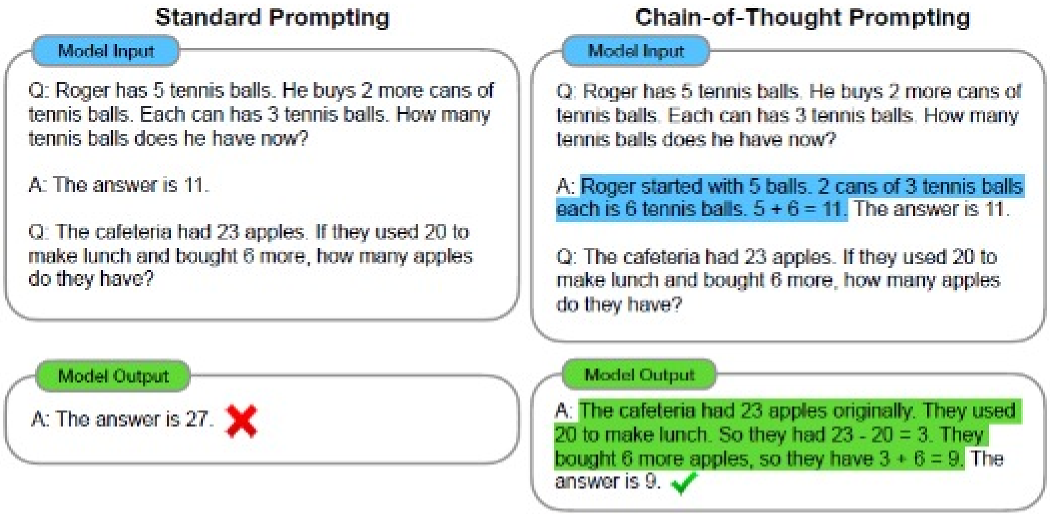


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

More Prompts:
(Tasks tested)

PROMPT FOR MATH WORD PROBLEMS
Q: 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?
A: 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.

PROMPT FOR LAST LETTER CONCATENATION
Q: Take the last letters of the words in "Elon Musk" and concatenate them.
A: The last letter of "Elon" is "n". The last letter of "Musk" is "k". Concatenating them is "nk". The answer is nk.

PROMPT FOR COIN FLIP

Q: A coin is heads up. Ka flips the coin. Sherrie flips the coin. Is the coin still heads up?

A: The coin was flipped by Ka and Sherrie. So the coin was flipped 2 times, which is an even number. The coin started heads up, so after an even number of flips, it will still be heads up. So the answer is yes.

PROMPT FOR CSQA (CommonSenseQA)

Q: What do people use to absorb extra ink from a fountain pen? Answer Choices: (a) shirt pocket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter

A: The answer must be an item that can absorb ink. Of the above choices, only blotters are used to absorb ink.

So the answer is (e).

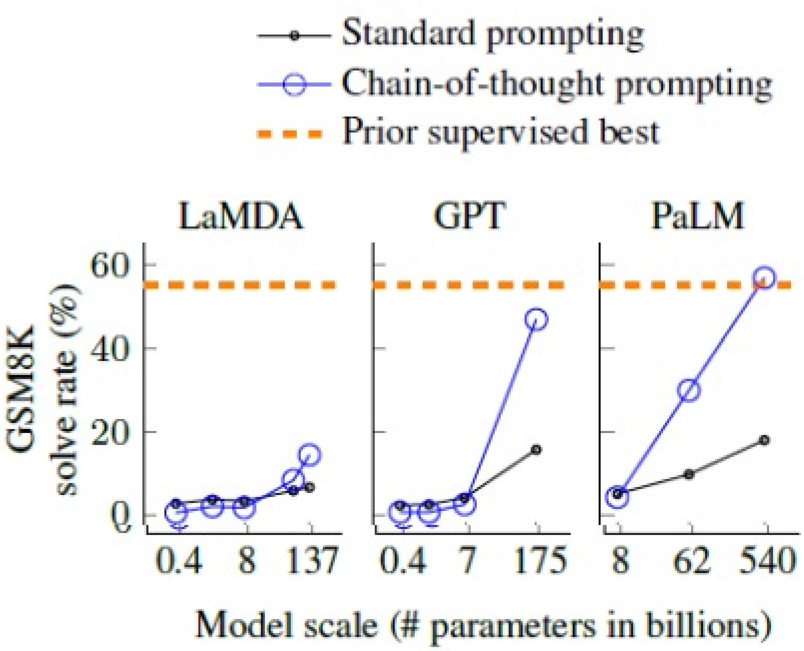
Benefits:
. decomposition of complex problems into intermediate steps, allowing for the allocation of additional computation to tasks that require more reasoning steps.
. interpretability by offering insights into the model's decision-making process.
. it can be utilized in any task that humans can solve through language-based reasoning.

Article contribution
Whereas commom approaches improve or augment the input part of the prompt (e.g., instructions that are prepended to inputs), CoT work takes the orthogonal direction of augmenting the outputs of language models with a chain of thought.

Results

Robustness of CoT
successful use of chain of thought does not depend on a particular linguistic style
Obs.: Sensitivity to exemplars is a key consideration of prompting approaches [varying the permutation of few-shot exemplars can cause the accuracy of GPT-3 on SST-2 to range from near chance (54.3%) to near state of the art (93.4%)]

CoT prompting does not positively impact performance for small models, and only yields performance gains when used with models of 100B parameters or more.
Models of smaller scale produced fluent but illogical chains of thought, leading to lower performance than standard prompting.



Basic doubt: Why LaMBDa models did not get well in the reusults?