Laura Perez-Beltrachini

Research Fellow

University of Edinburgh

UKRI Innovate UK BridgeAI

Scottish AI Alliance

Addressing Hallucinations in Generative AI for Customer Service Applications

Innovate UK 10092998 / Collaborative R&D

Accelerating trustworthy AI

Feb 2024 - Feb 2025

Innovate UK BridgeAI Regional Roadshow Series

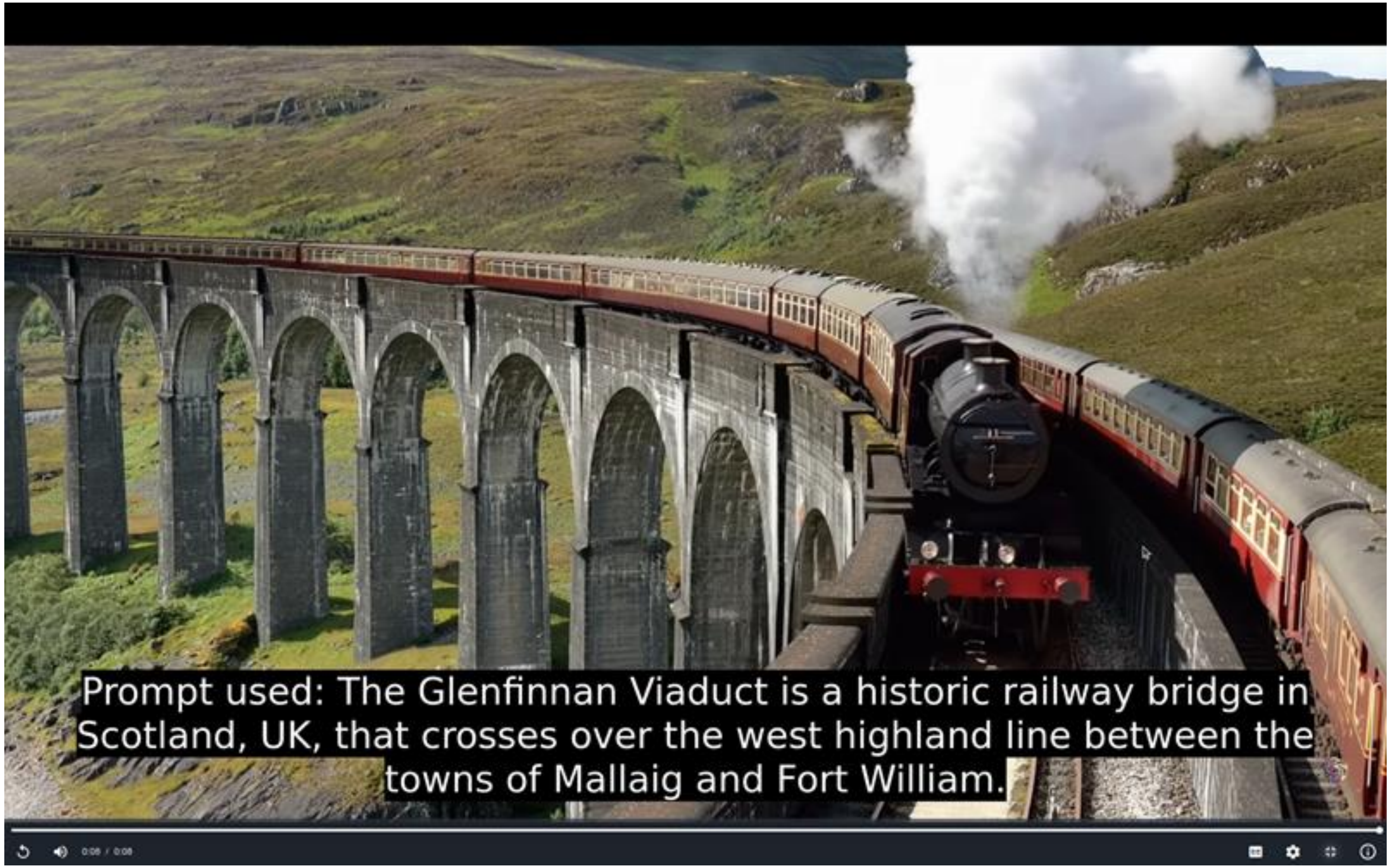
12th May 2025

Laura Perez-Beltrachini

THE UNIVERSITY of EDINBURGH

Edinburgh NLP
University of Edinburgh
Natural Language Processing

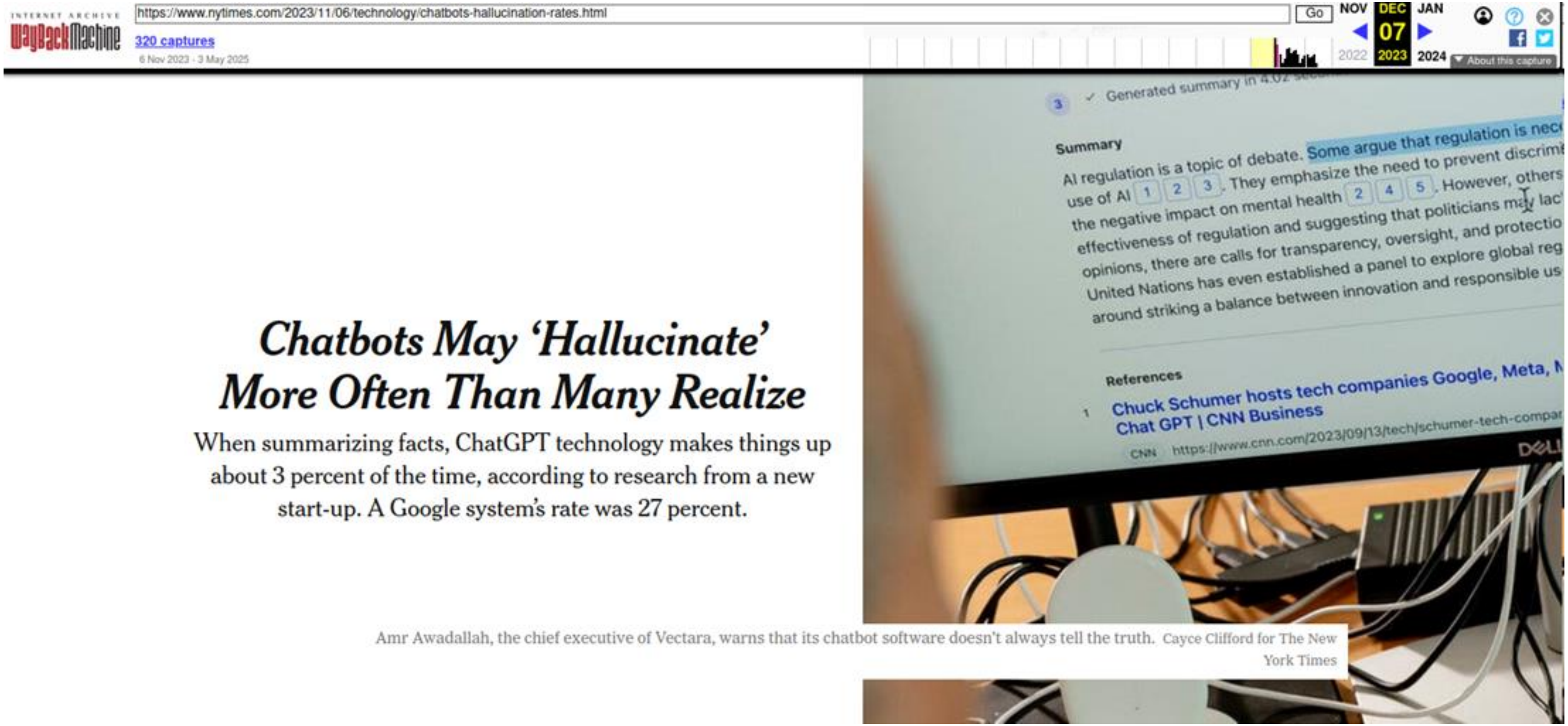
Hallucinations in Generative AI



Prompt used: The Glenfinnan Viaduct is a historic railway bridge in Scotland, UK, that crosses over the west highland line between the towns of Mallaig and Fort William.

Generated by Sora (OpenAI's text-to-video model)
https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)

Hallucinations in Large Language Models (LLMs)



INTERNET ARCHIVE
WayBackMachine
Go
NOV DEC JAN
07
2022 2023 2024
About this capture
3 ✓ Generated summary in 4.02 seco
Summary
AI regulation is a topic of debate. Some argue that regulation is nec
use of AI 1 2 3 . They emphasize the need to prevent discrimi
the negative impact on mental health 2 4 5 . However, others
effectiveness of regulation and suggesting that politicians may lac
opinions, there are calls for transparency, oversight, and protectio
United Nations has even established a panel to explore global reg
around striking a balance between innovation and responsible us
References
1 Chuck Schumer hosts tech companies Google, Meta, N
Chat GPT | CNN Business
CNN https://www.cnn.com/2023/09/13/tech/schumer-tech-compa
DELL

https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html

320 captures

6 Nov 2023 - 3 May 2025

Chatbots May 'Hallucinate' More Often Than Many Realize

When summarizing facts, ChatGPT technology makes things up about 3 percent of the time, according to research from a new start-up. A Google system's rate was 27 percent.

Amr Awadallah, the chief executive of Vectara, warns that its chatbot software doesn't always tell the truth. Cayce Clifford for The New York Times

R&D to Address Hallucinations in LLMs for Customer Service Chatbots

Large company	NatWest Algomo	banking dataset, piloting solution, and providing customer access
SME	yu life THE UNIVERSITY of EDINBURGH (Lead Participant) University of Essex	core conversational AI platform and productization. technical implementation and end-user of the solution
SME		

We thank Harry Sfyrakis (Algomo Ltd) for this slide.

Our Research on Detecting Hallucinations

generation
fact-checking

Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks (Zhang, Xu, Perez-Beltrachini, EACL 2024)

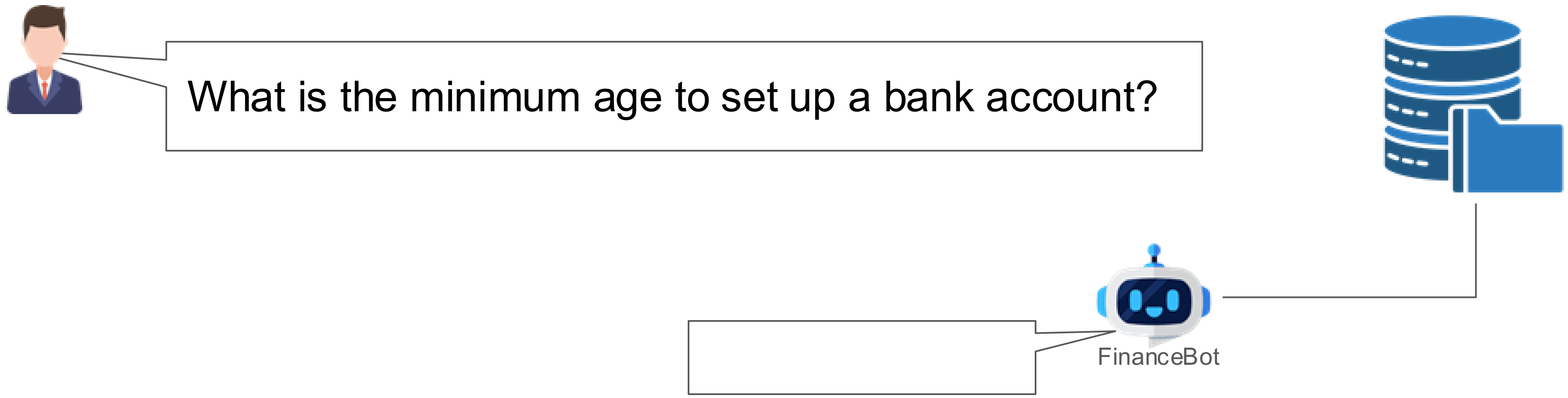
Leveraging entailment judgements in cross-lingual summarisation (Zhang and Perez-Beltrachini, ACL 2024)

model
uncertainty

Uncertainty Quantification in Retrieval Augmented Question Answering (Perez-Beltrachini and Lapata, Under Submission)

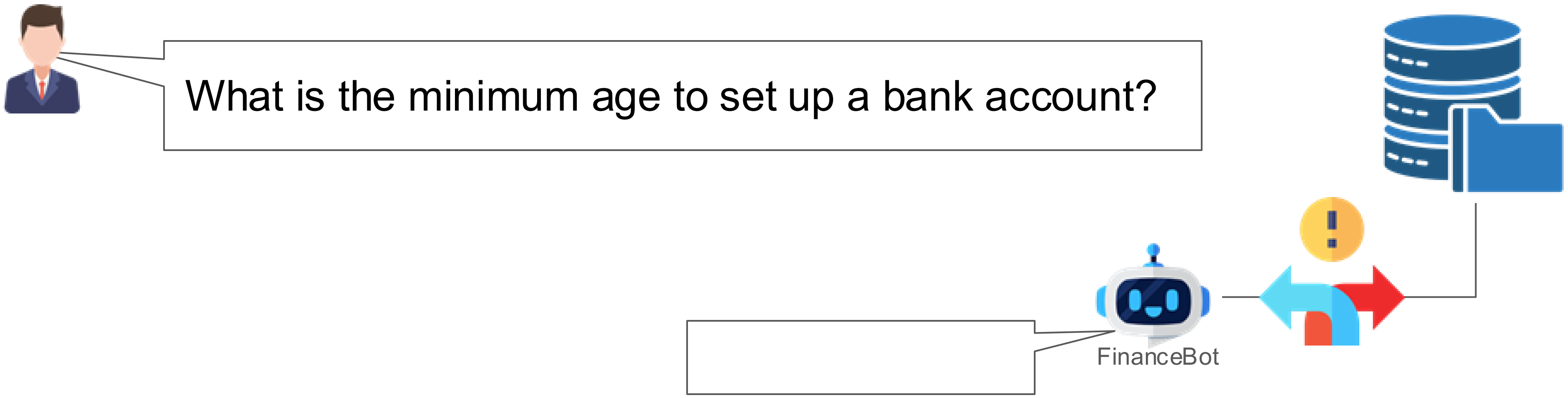
THE UNIVERSITY
of EDINBURGH

Retrieval Augmented Question Answering



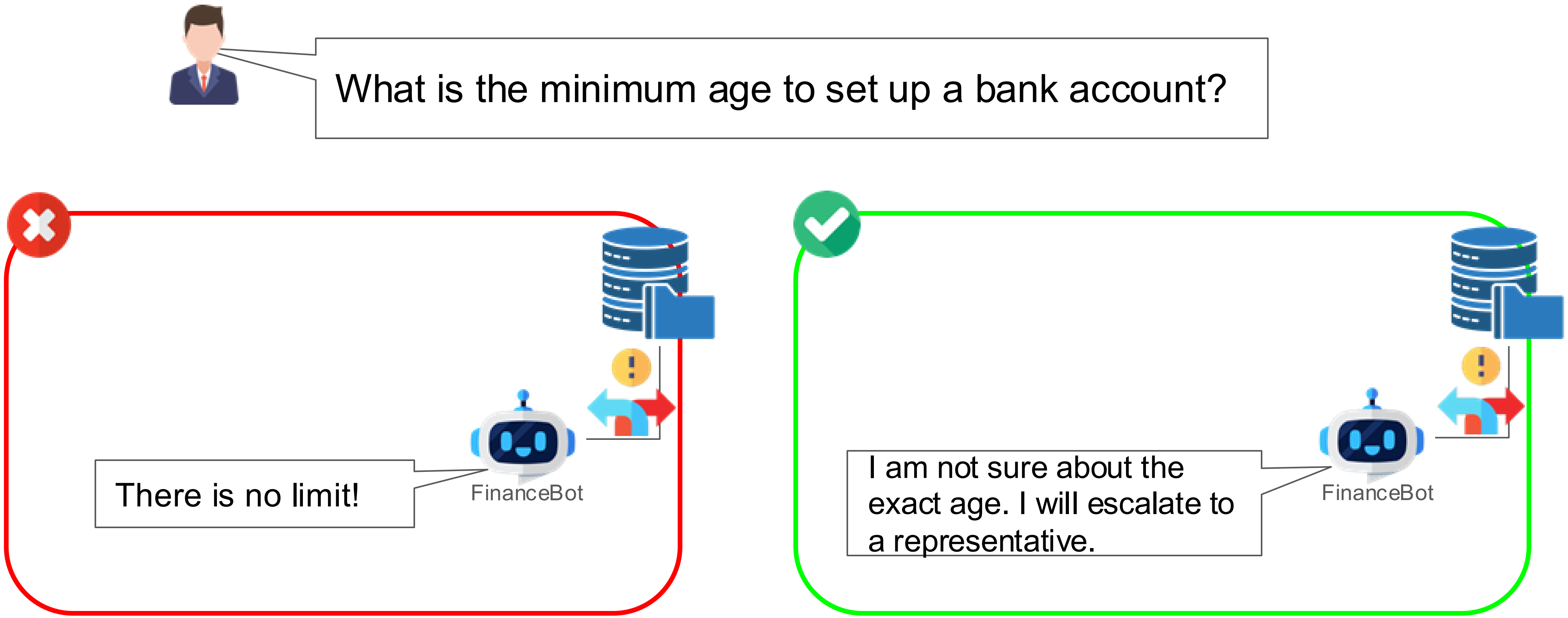
What is the minimum age to set up a bank account?

Retrieval Augmented Question Answering



What is the minimum age to set up a bank account?

Hallucinations and Answer Uncertainty



What is the minimum age to set up a bank account?

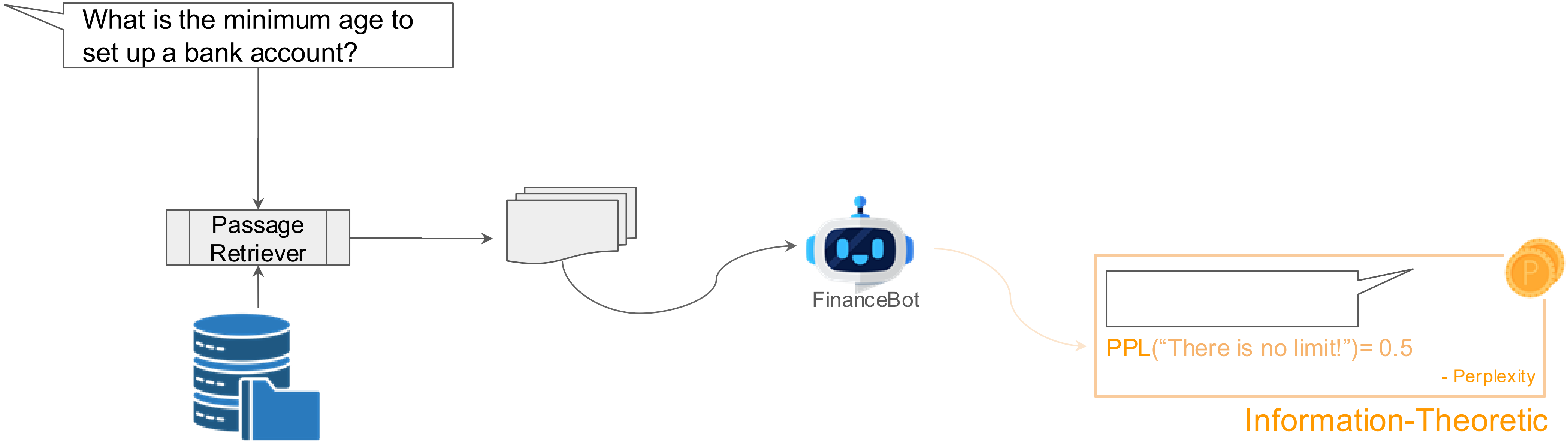
There is no limit!

I am not sure about the exact age. I will escalate to a representative.

Hallucinated Answer

Safe Answer

Answer Uncertainty Estimation?



What is the minimum age to set up a bank account?

Passage Retriever

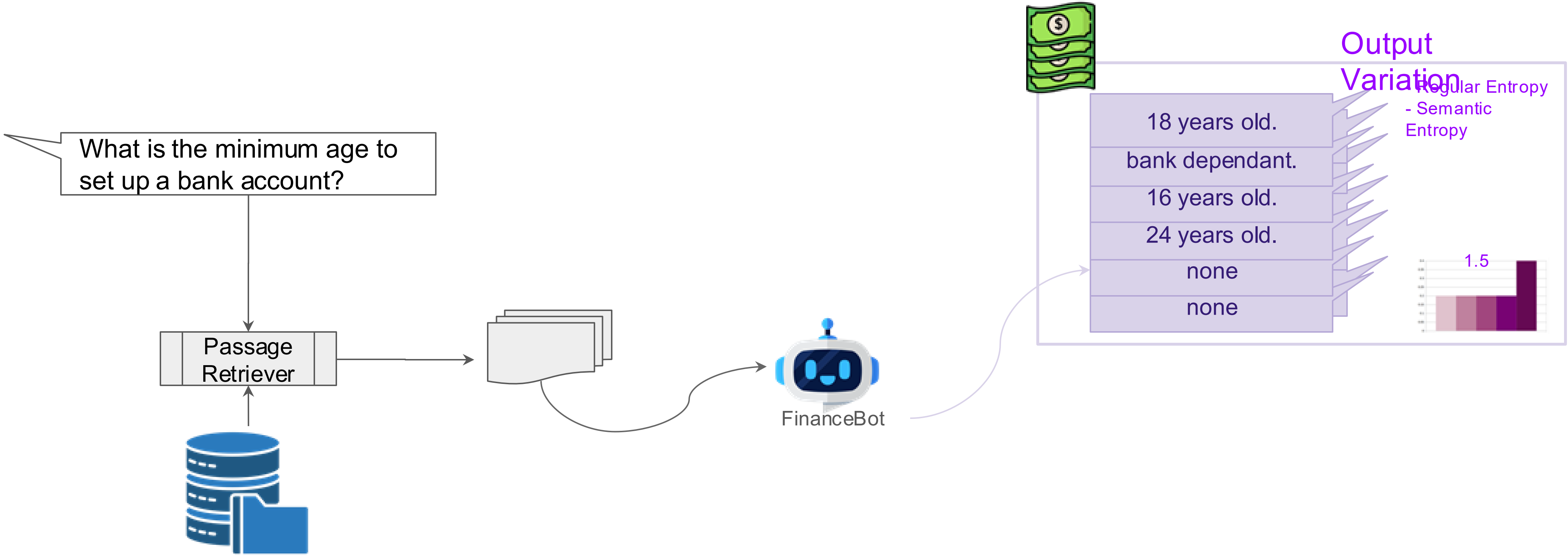
FinanceBot

PPL(“There is no limit!”)= 0.5

- Perplexity

Information-Theoretic

Answer Uncertainty Estimation?



What is the minimum age to set up a bank account?

Passage Retriever

FinanceBot

Output Variation

- Regular Entropy
- Semantic Entropy

18 years old.

bank dependant.

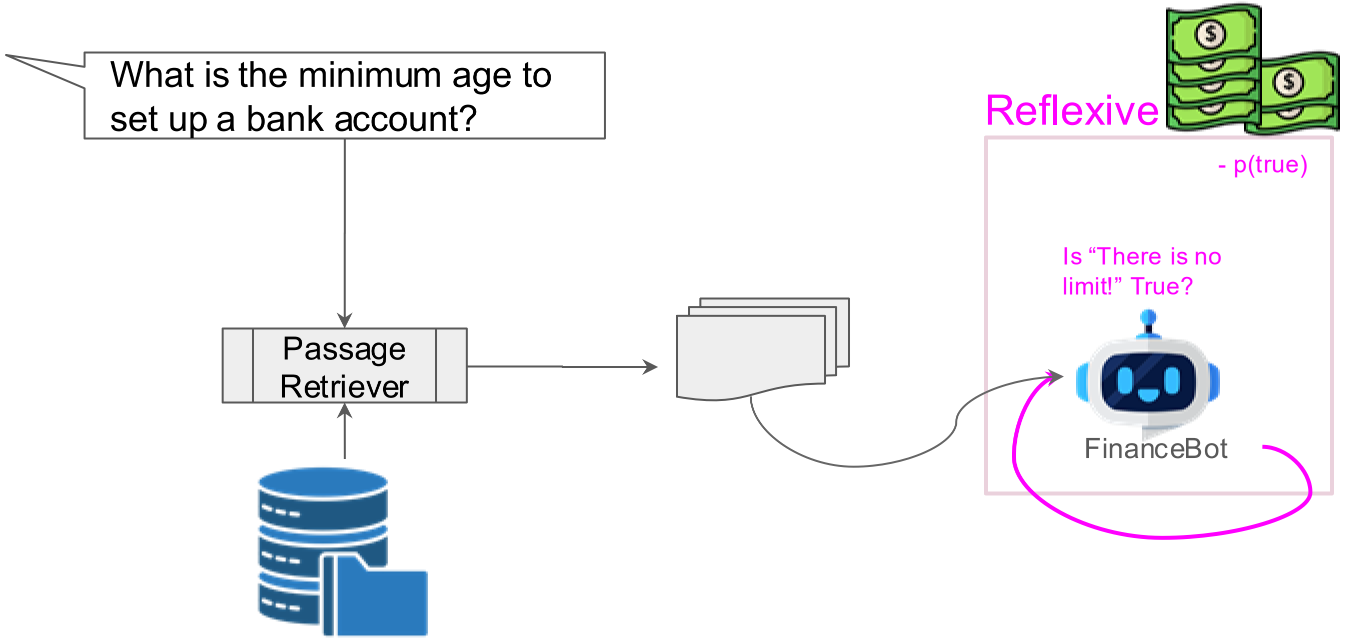
16 years old.

24 years old.

none

none

Answer Uncertainty Estimation?



What is the minimum age to set up a bank account?

Passage Retriever

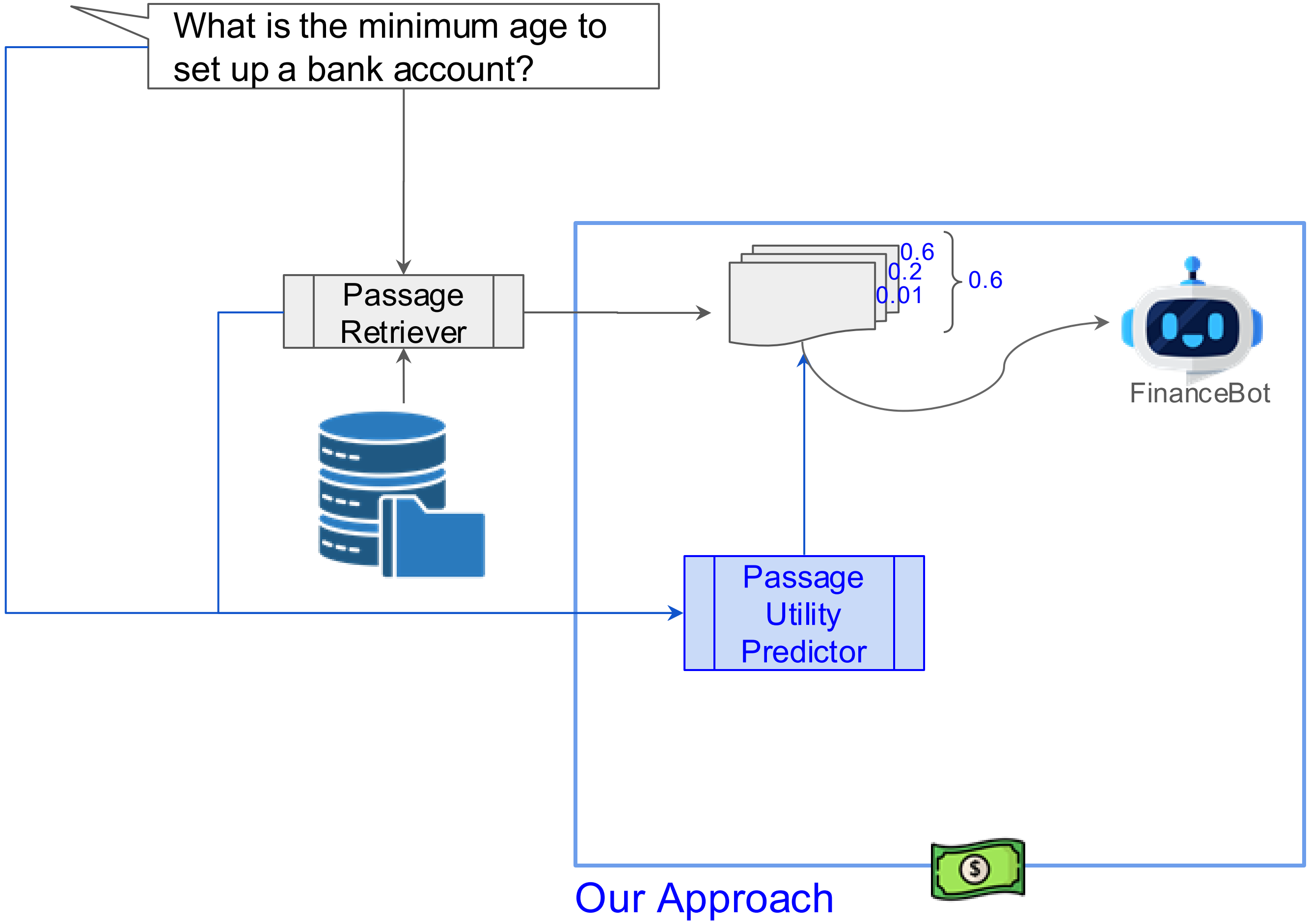
Reflexive

- p(true)

Is “There is no limit!” True?

FinanceBot

Answer Uncertainty via Passage Utility

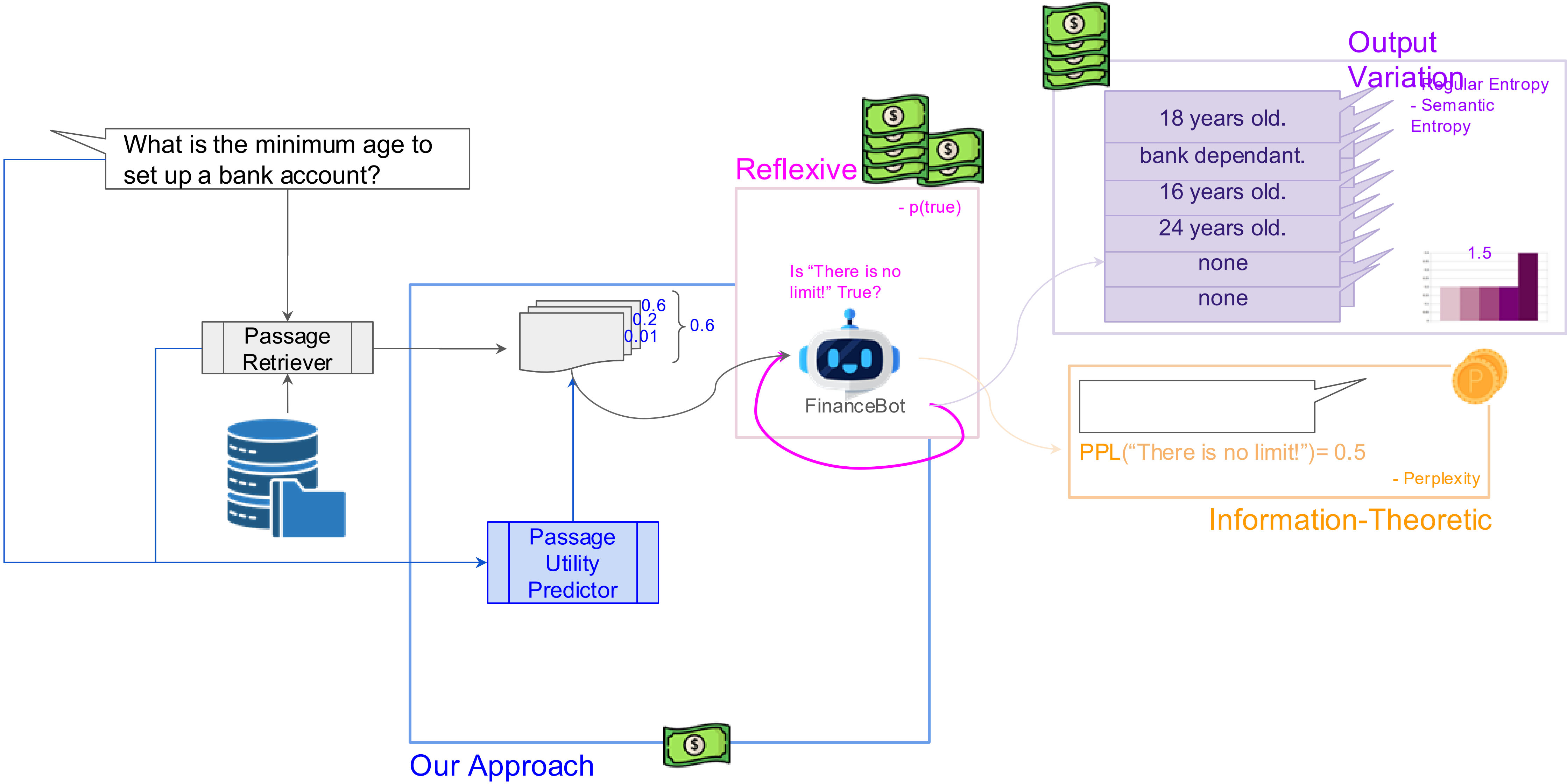

What is the minimum age to set up a bank account?
Passage Retriever
0.6
0.2
0.01
0.6
FinanceBot
Passage Utility Predictor
Our Approach

1° observe QA performance on individual passages and target LLM(θ), Passage Utility.

2° train a secondary small model to predict Passage Utility.

3° estimate Answer Uncertainty from Passage Utilities.

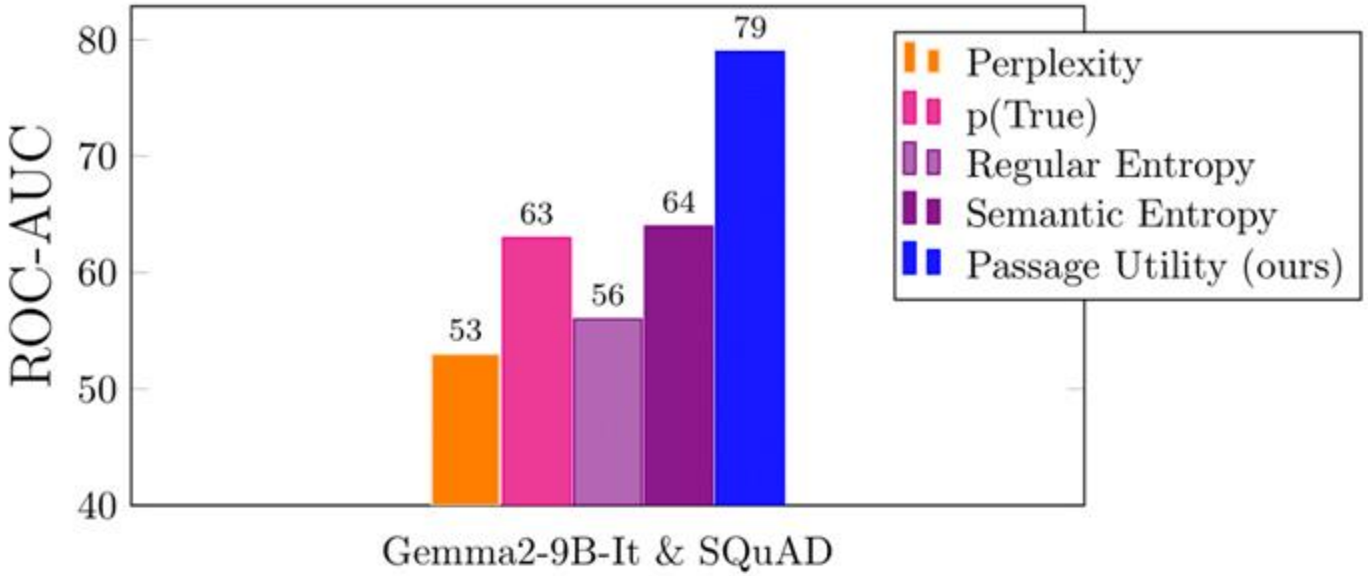
Answer Uncertainty Estimation?



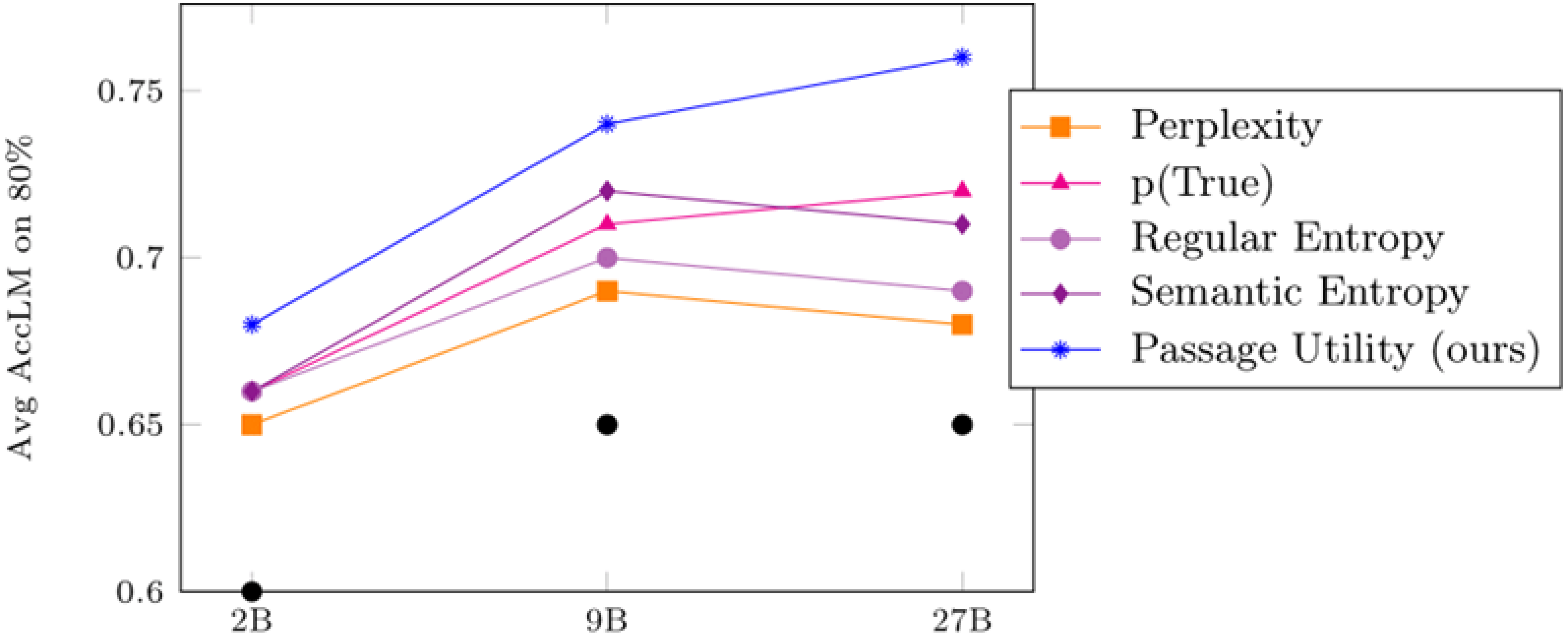
Successful and Lightweight Hallucination Detection

Natural Questions, TriviaQA, WebQuestions, SQuAD, PopQA, RefuNQ

OpenDomainBot Llama-3.1-8B, Gemma2-9B, Mistral-v0.3-7B, Gemma2-2B, Gemma2-27B



Improved Accuracy via Selective Answering

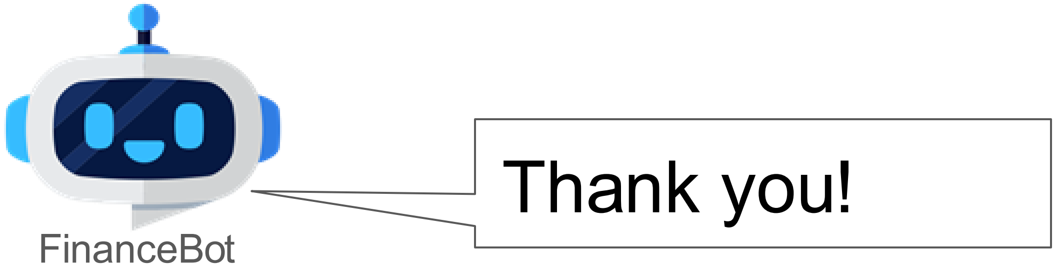


X-axes Gemma2 -2B, -9B, -27B.
Average AccLM (LLM- based accuracy) is across the six QA tasks.
Black dots indicate accuracy when always answering.

Going Forward

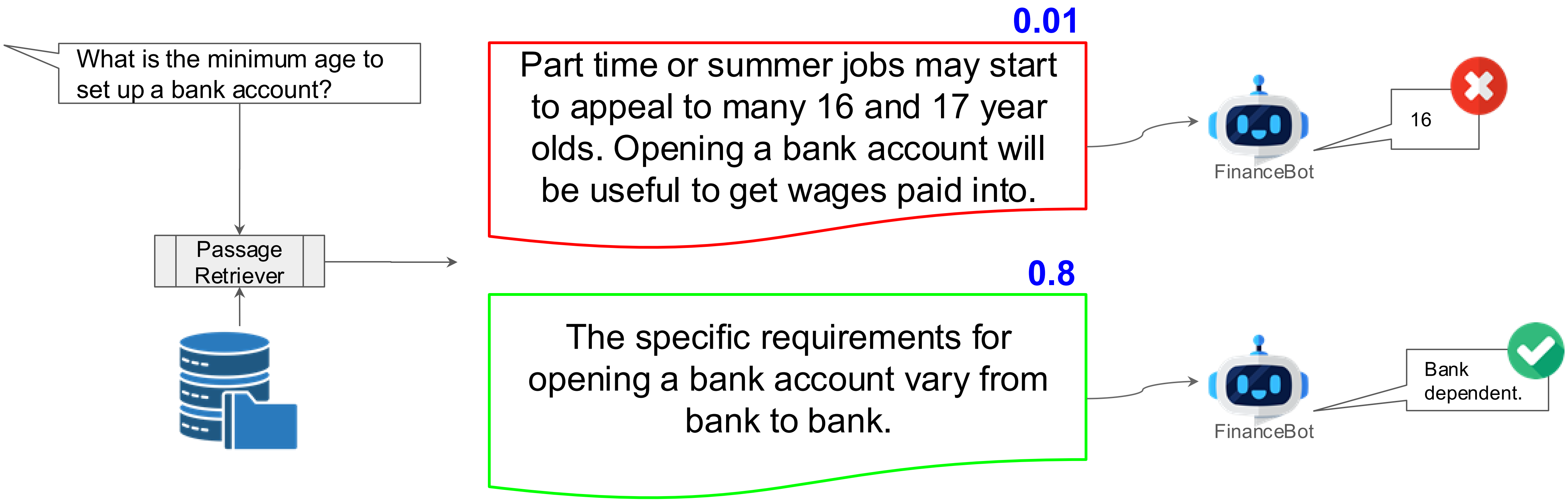
- Generalisation of the Passage Utility predictor:
  - Unseen QA tasks and LLMs,
  - Multi-Hop reasoning,
  - Long-form and subjective answers.

- Answering ambiguous questions.



We thank Roundicons, Luch Phou, littleicon, Freepik, Aranagraphics, HAJICON, Good Ware from the Flaticon library.

Observed Low/High Passage Utility



What is the minimum age to set up a bank account?

Passage Retriever

0.01

Part time or summer jobs may start to appeal to many 16 and 17 year olds. Opening a bank account will be useful to get wages paid into.

FinanceBot

16

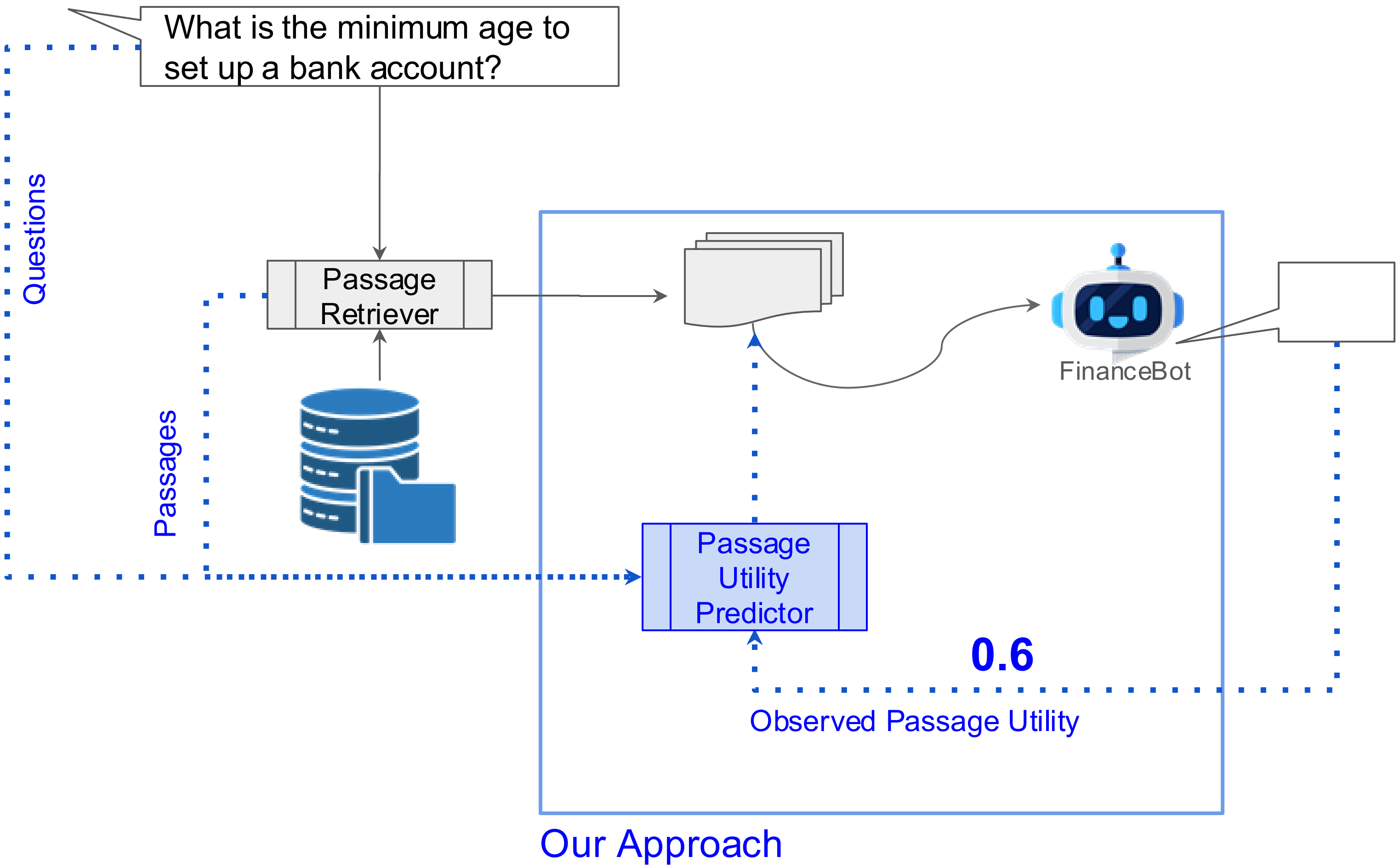
0.8

The specific requirements for opening a bank account vary from bank to bank.

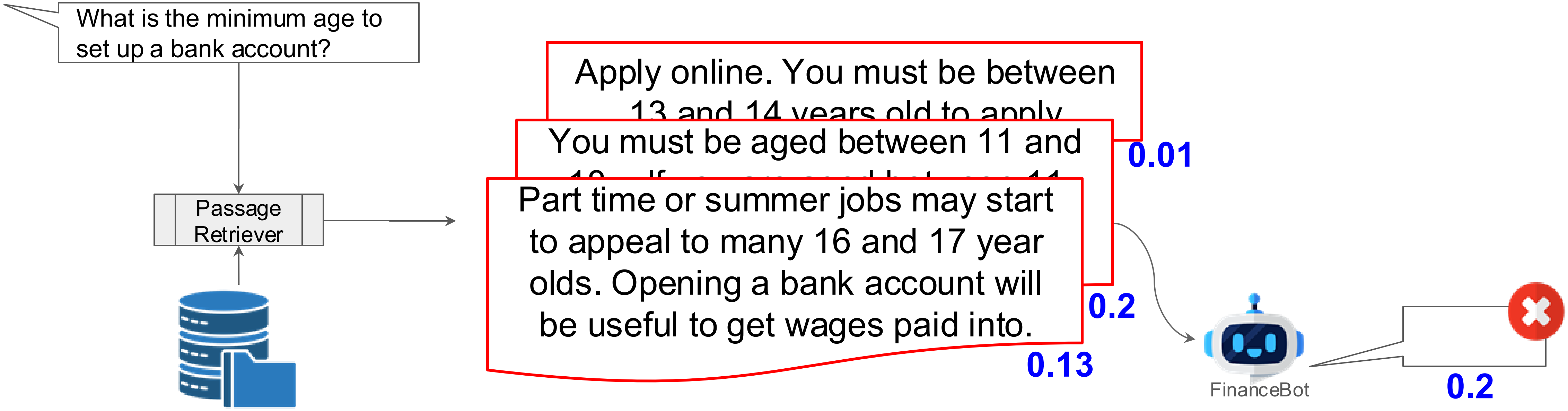
FinanceBot

Bank dependent.

Train a Passage Utility Predictor



Answer Uncertainty via Passage Utility



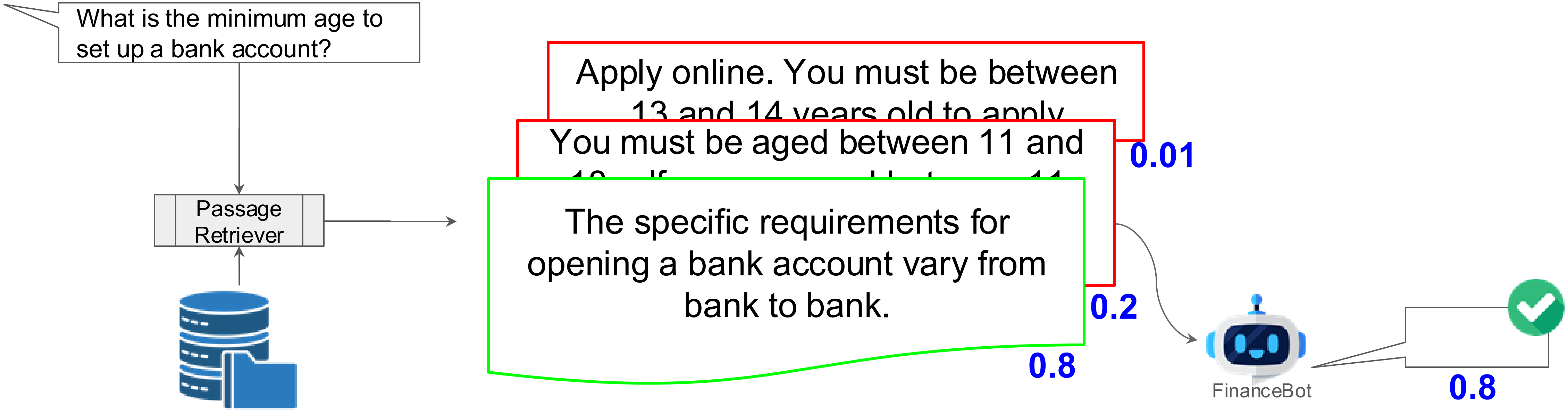
What is the minimum age to set up a bank account?

Apply online. You must be between 13 and 14 years old to apply

You must be aged between 11 and

Part time or summer jobs may start to appeal to many 16 and 17 year olds. Opening a bank account will be useful to get wages paid into.

Answer Uncertainty via Passage Utility



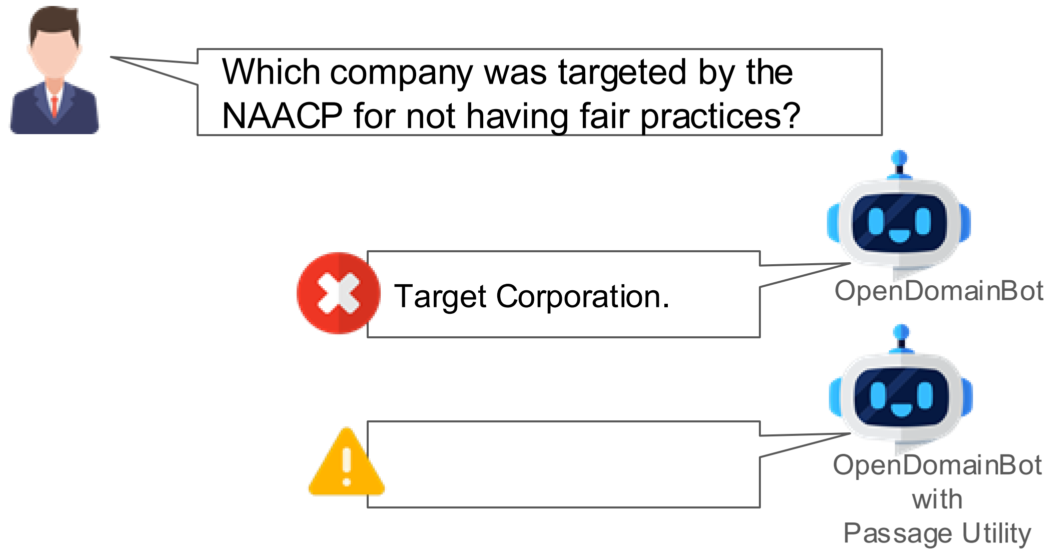
What is the minimum age to set up a bank account?

Apply online. You must be between 13 and 14 years old to apply.

You must be aged between 11 and 12. If you are aged between 11

The specific requirements for opening a bank account vary from bank to bank.

Prediction of Hallucination
Llama3.1-8B & question from SQuAD



Which company was targeted by the NAACP for not having fair practices?

Target Corporation.

Target Corporation. of Colored People has repeatedly given Target failing grades on its annual Economic Reciprocity Initiative report card, a measure of the company's ``commitment to the African-American citizenry". In 2003 and 2005, the NAACP has rated Target an ``F" on this report; in 2004, Target was rated a ``D-". In 2006, when Target was asked why it didn't participate in the survey again, a representative explained, ``Target views diversity as being inclusive of all people from all different backgrounds, not just one group." In February 2006, the National Federation of the Blind (NFB) filed a class action [cont.]

Utility: -3.59

Misleading

Reactions to the Duke lacrosse case. formed an opinion on the case. North Carolina NAACP Legal Redress Chair, Al McSurely, explained that ``The NAACP stands for fair play for all parties, zealous investigation and deep concern for the survivors of racist/sexist attacks." At the same time, some have criticized the NAACP for making statements that portrayed the players as racist despite evidence to the contrary, using the case to promote the group's cause, and implying guilt. McSurely stated that ``[w]ithin five minutes, the men threatened the women with racial and misogynist verbal assaults, [cont.]

Utility: -3.67

Loosely Related

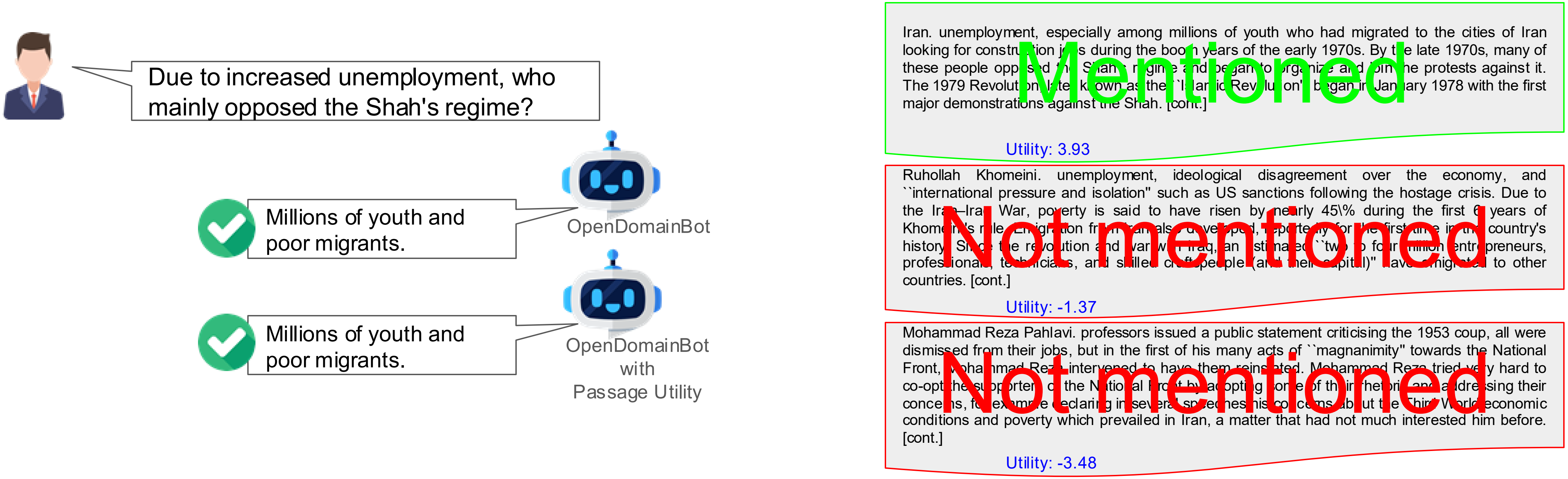
Philadelphia transit strike of 1944. (PRTEU), Frank Carney, proved to be equally reticent and claimed that he was not authorized by the union members to consider a request to allow promotions of black employees. The black PTC employees enlisted the help of the NAACP and started lobbying the federal authorities, particularly the Fair Employment Practices Commission (FEPC), to intervene. The Fair Employment Practices Commission, created by an executive order of the President in 1941, was charged with ensuring non-discrimination employment practices by government contractors. [cont.]

Utility: -3.49

Not a company

The gold standard answer is “Lockheed Aircraft Corporation.”

Prediction of Correct Answer
Llama3.1-8B & question from SQuAD


Due to increased unemployment, who mainly opposed the Shah's regime?
Millions of youth and poor migrants.
OpenDomainBot
Millions of youth and poor migrants.
OpenDomainBot with Passage Utility
Mentioned
Not mentioned
Not mentioned

Iran. unemployment, especially among millions of youth who had migrated to the cities of Iran looking for construction jobs during the boom years of the early 1970s. By the late 1970s, many of these people opposed the Shah's regime and began to organize and join the protests against it. The 1979 Revolution, later known as the ``Islamic Revolution'', began in January 1978 with the first major demonstrations against the Shah. [cont.]

Utility: 3.93

Ruhollah Khomeini. unemployment, ideological disagreement over the economy, and ``international pressure and isolation" such as US sanctions following the hostage crisis. Due to the Iran–Iraq War, poverty is said to have risen by nearly 45\% during the first 6 years of Khomeini's rule. Emigration from Iran also developed, reportedly for the first time in the country's history. Since the revolution and war with Iraq, an estimated ``two to four million entrepreneurs, professionals, technicians, and skilled craftspeople (and their capital)" have emigrated to other countries. [cont.]

Utility: -1.37

Mohammad Reza Pahlavi. professors issued a public statement criticising the 1953 coup, all were dismissed from their jobs, but in the first of his many acts of ``magnanimity" towards the National Front, Mohammad Reza intervened to have them reinstated. Mohammad Reza tried very hard to co-opt the supporters of the National Front by adopting some of their rhetoric and addressing their concerns, for example declaring in several speeches his concerns about the Third World economic conditions and poverty which prevailed in Iran, a matter that had not much interested him before. [cont.]

Utility: -3.48

The gold standard answer is “Millions of youth who had migrated to the cities.”

Andrew Low

COO

FC Laboratories

UKRI Innovate UK BridgeAI

Scottish AI Alliance

brainfit™

powered by FCLABS

Andy Low
andy@fclabs.co.uk
COO
brainfitfirst.com