



Temporal Validity Change Prediction

Georg Wenzel (Supervisor: Adam Jatowt)


The University of Innsbruck was founded in 1669 and is one of Austria's oldest universities. Today, with over 28.000 students and 5.000 staff, it is western Austria's largest institution of higher education and research. **For further information visit: www.uibk.ac.at**

Language Models




Reasoning in Language Models

How many times does the letter "r" occur in the word strawberry?

 The letter "r" occurs twice in the word "strawberry."


Can you highlight the letter r in the word?

 Sure, here's the word with the letter "r" highlighted:


strawbrry

Temporal Commonsense Reasoning

“Barack Obama is the President of the United States of America”



Duration
4 years



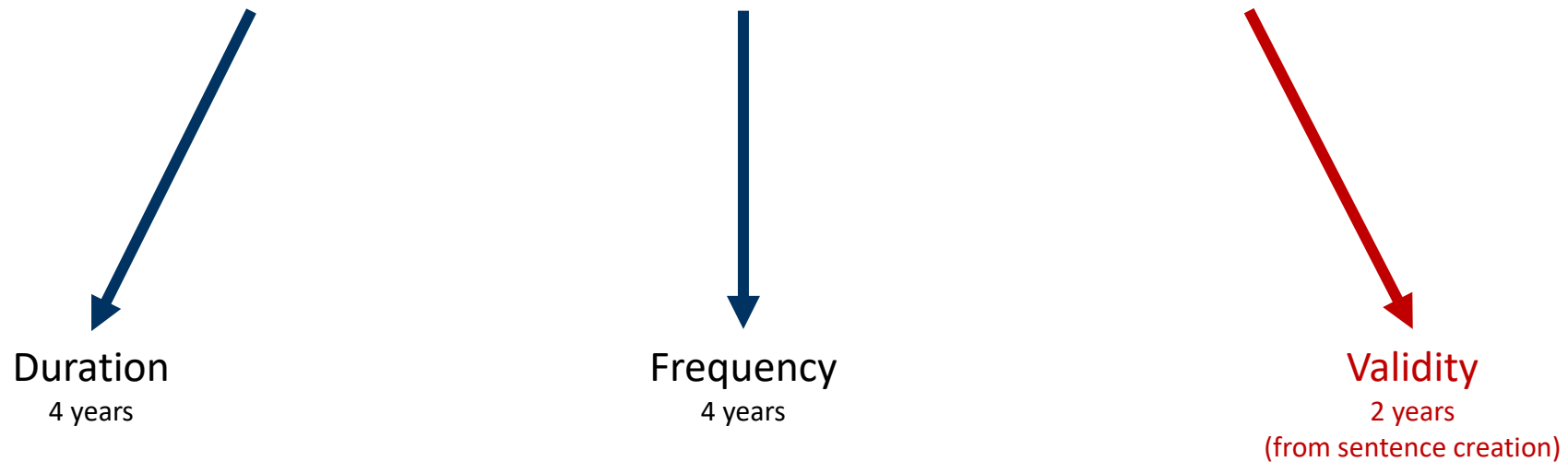
Frequency
4 years



Validity
2008-2016

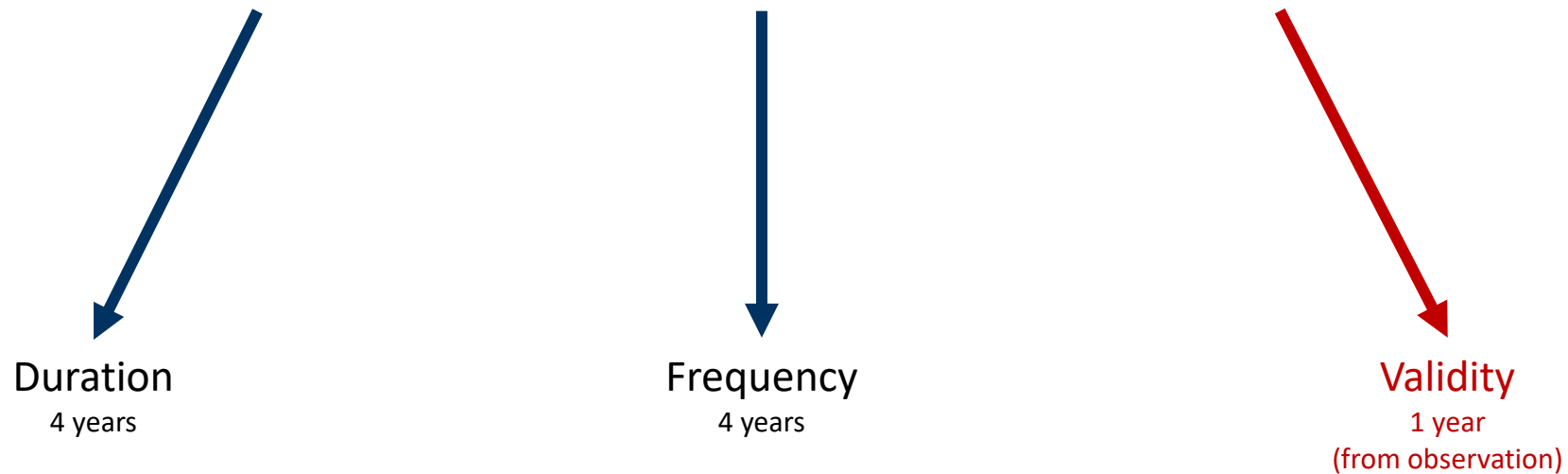
Temporal Commonsense Reasoning

“Barack Obama is the President of the United States of America”
“He took office two years ago”

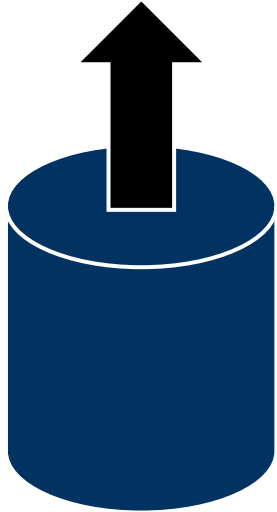


Temporal Commonsense Reasoning

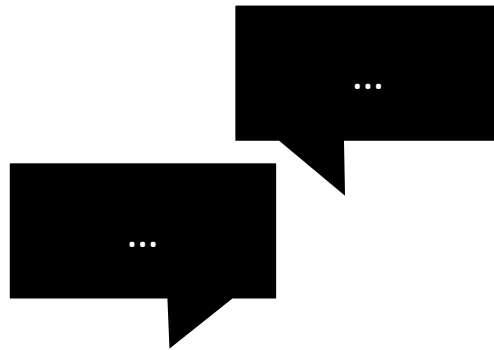
“Barack Obama is the President of the United States of America”
“He took office two years ago”
(Posted one year ago)



Use Cases of Temporal Validity



Information Retrieval

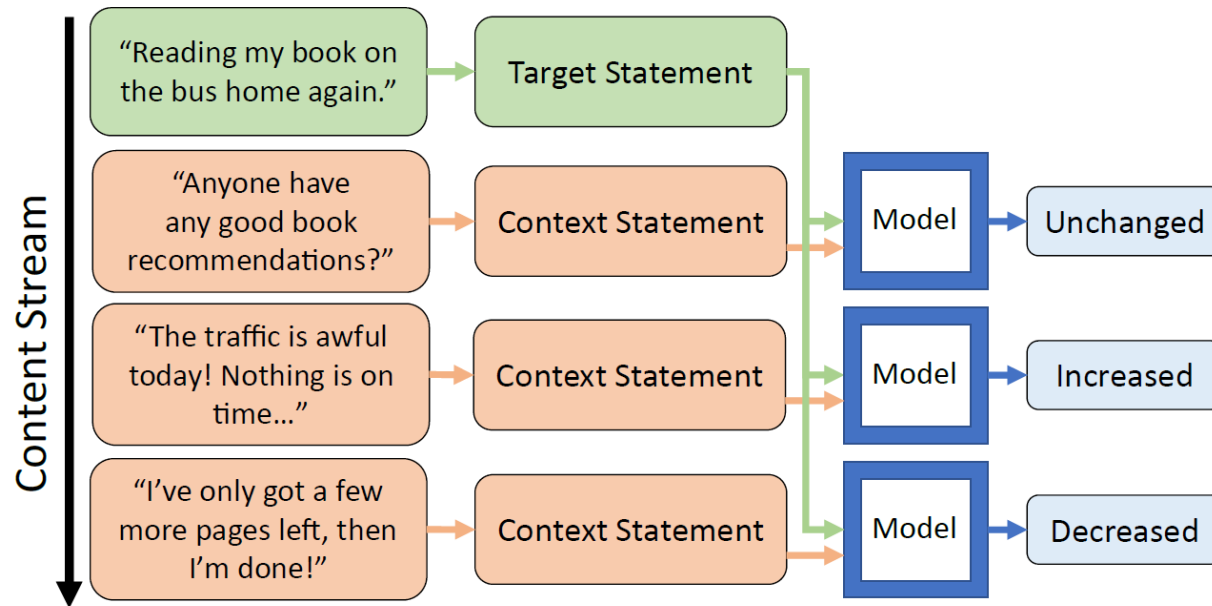


Tracking Contemporaneity

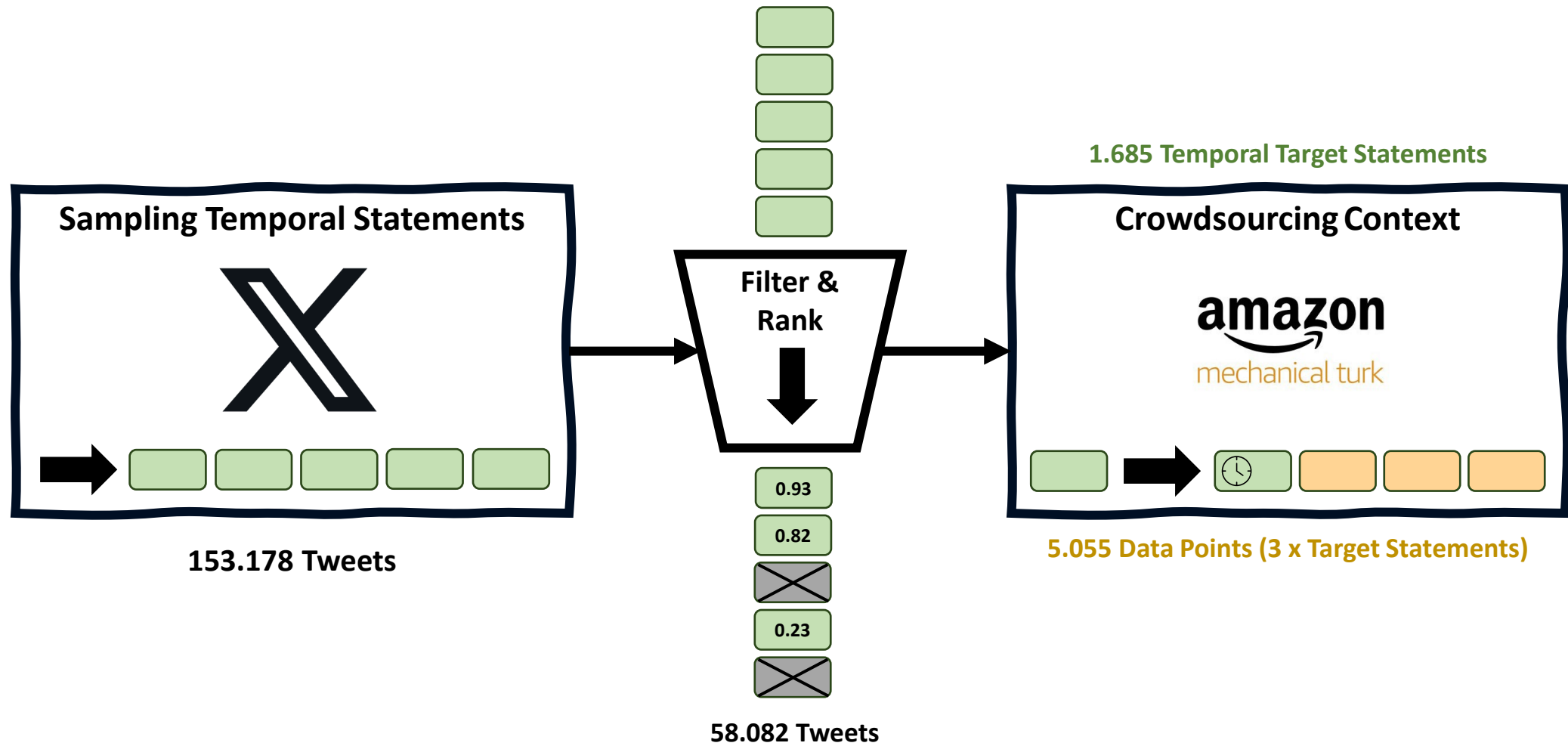


Content Prioritization

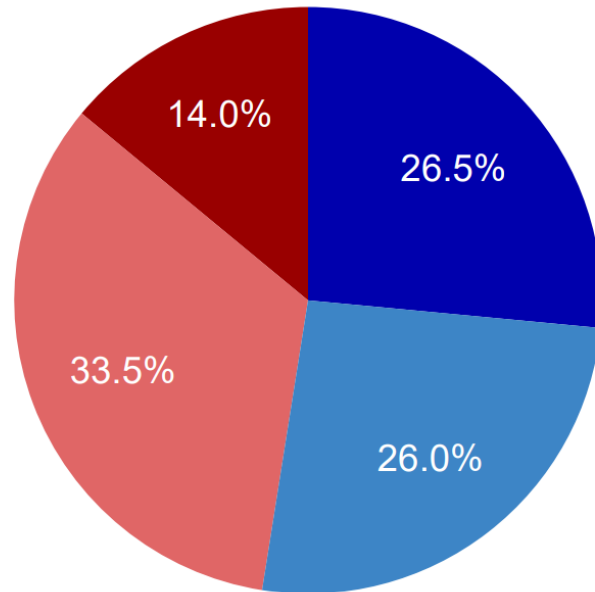
Temporal Validity Change Prediction (TVCP)



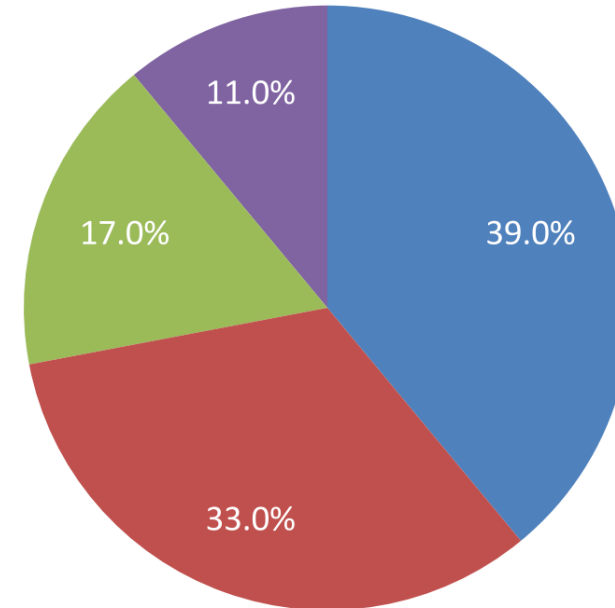
Dataset Creation



Qualitative Analysis of Dataset



- Context alters expected occurrence time explicitly
- Context alters expected duration explicitly
- Context allows for refined duration estimate
- Context allows for refined occurrence time estimate



- Action
- Multiple
- Temporary State
- Event

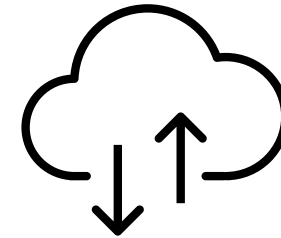
Evaluated Language Models



Fine-tuned (LM)

(BERT, RoBERTa, SelfExplain)

- Millions of parameters
- Fine-tuned, specialized model
- Calculate output neurons for input
- Receive a numeric output (vector)
- Neurons correspond to classes




Few-shot prompted (LLM)

(Mixtral-8x7B, Llama 2, GPT-3.5, GPT-4)

- (Likely) billions of parameters
- Pre-trained, generalized model
- Prompted via API
- Receive a textual response
- Parse class from the response

Fine-Tuned Models Outperform LLMs

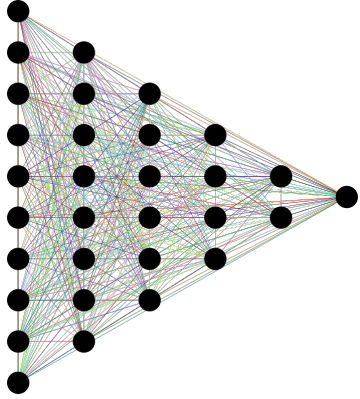
GPT-3.5



EM: 31.1%
Few-shot prompted



SelfExplain



EM: 69.8%
Fine-tuned

LLM Performance Stagnates

GPT-4



EM: 30.4%
\$10.00 / 1M tokens



GPT-3.5



EM: 31.1%
\$0.50 / 1M tokens

Prompt Engineering Improves LLMs

GPT-3.5 (Thesis)



EM: 29.3%

- Does not know the underlying class hierarchy
- Limited chain-of-thought reasoning
- Three few-shot samples (one per class)



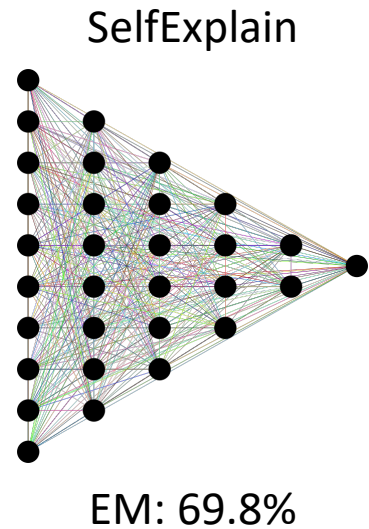
GPT-3.5 (ACL)



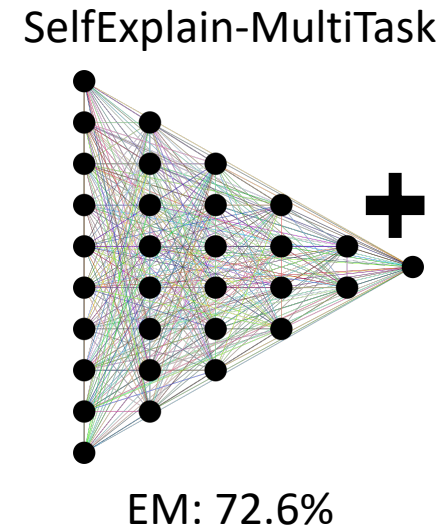
EM: 31.1%

- Knows and predicts the underlying class hierarchy
- Longer chain-of-thought reasoning process
- Nine perturbed few-shot samples (three per class)

Multitask Learning Improves Fine-Tuned Models



- Only trained on ternary output label classification



- Trained on ternary output label classification **and** related temporal validity tasks

