Sequential Text-based Knowledge Update with Self-Supervised Learning for Generative Language Models

Hao-Ru Sung[1], Ying-Jhe Tang[2], Yu-Chung Cheng[1],
Pai-Lin Chen[1], Tsai-Yen Li[1], Hen-Hsen Huang[2]

[1]National Chengchi University
[2]Academia Sinica

Introduction - Background and Motivation

- In a ever-changing world, knowledge continuously evolves, requiring ongoing updates to remain relevant and accurate. (e.g. books, web pages, etc.)
- Keeping knowledge up-to-date is a complex and time-consuming task.
  - Large knowledge carriers like Wikipedia
    - Large team of editors working to keep its 6.5 million articles up-to-date.
  - Online News Platform
    - Manual writing and editing of updates

WIKIPEDIA

Introduction - Background and Motivation

- Maintaining current knowledge is an ongoing process, particularly for unfolding events .
- Example: Natural Disasters like Floods
  - Initial reports give basic overview
  - Continuous updates refine understanding.
  - Additional details added
  - Inaccuracies corrected.

Introduction - Goal

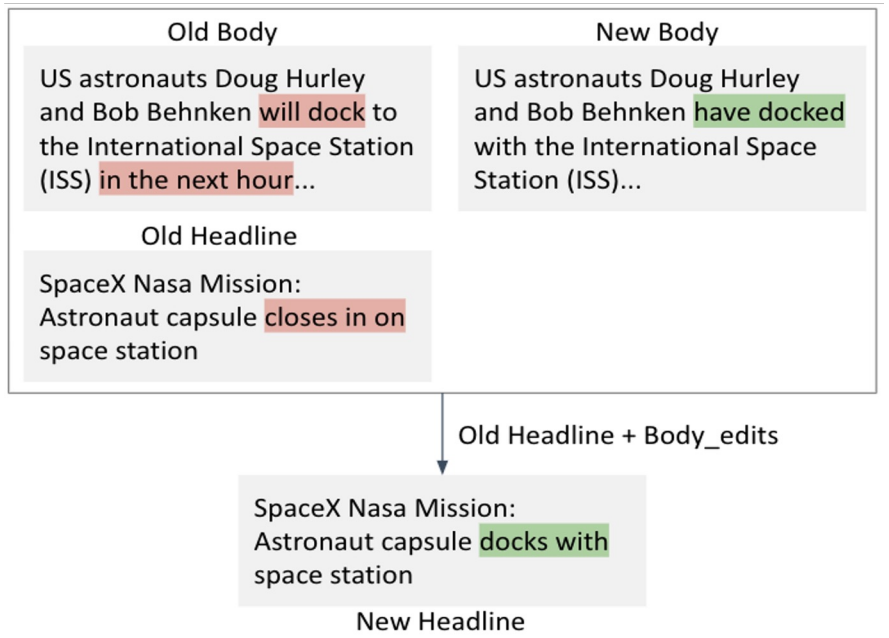
- **Objective:** Addressing the problem of updating knowledge of multiple consecutive texts through natural language.

- Our study focuses on natural language for knowledge representation.
- Introducing a hybrid learning method combined with self-supervised training.
- Outperformed ChatGPT in text-based sequential knowledge updates.

Related Works

Updated Headline Generation: Creating Updated Summaries for Evolving News Stories *[Panthaplackel et al., ACL 2022]*

For two different versions of News Stories, using the old headline and identifying the changes between the old body and the new body to generate a new headline.



Our study:

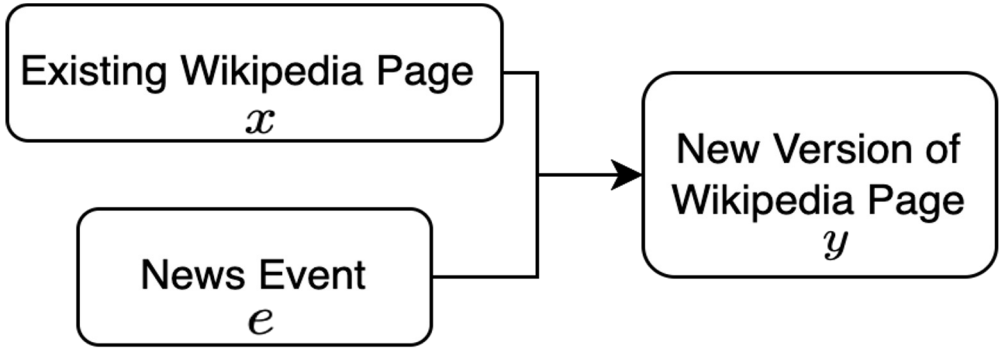
1. Focusing **on updating multiple sequential articles**
2. **Considers of Wikipedia summary without labelling the edited section.**
3. **Generates text similar in length to the original content.**

Related Works

A Multi-Grained Dataset for News Event Triggered Knowledge Update

[Lee et al., CIKM 2022]

- Single-round updates for multi-grained textual data.
- Preliminary results were presented using a baseline model.
- Focusing on short-term event updates



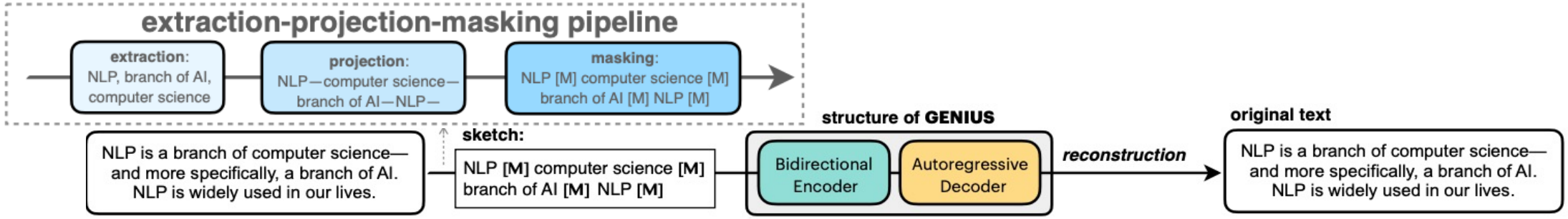
Our Study:

1. **Focusing on multiple rounds of updates**
2. **Using multiple training strategies**
3. **Focusing on long-term event updates**

Related Works

Sketch-based Language Model Pre-training via Extreme and Selective Masking for Text Generation and Augmentation

[Guo et al., 2022]



- To build a conditional language model to simulate this process, **generating a full text based on the sketch.**

Text-to-Sketch method:

1. **Extraction:** Use YAKE to find keywords, roughly 20% of the text.
2. **Projection:** Project these key portions back into the original text, allowing for overlaps.
3. **Masking:** Replace the remaining text with a mask token, about 73% on average.

Problem Formulation

$$x_t = \arg\max_y Pr(y|x_1, x_2, ..., x_{t-1}, e_t) \quad (1)$$

$$\approx \arg\max_y Pr(y|x_{t-1}, e_t) \quad (2)$$

Inputs

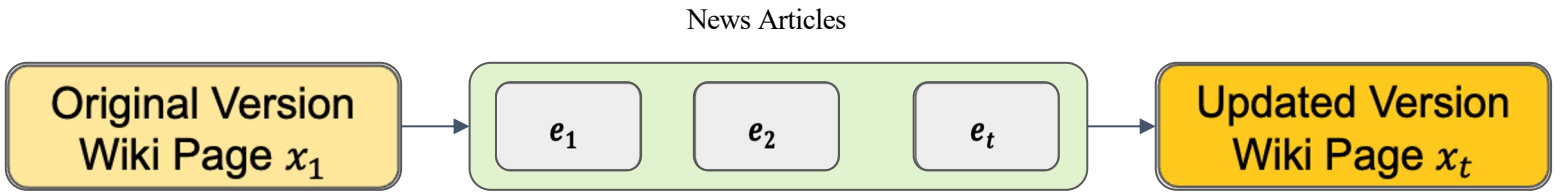
- A sequence **of text-based knowledge facts** until time t-1: **($x_1, x_2, ..., x_{t-1}$)**
- A piece of information about a **related event** that occurs at time t : e_t

Output

- The goal is to **generate** x_t, updating x_{t-1} based on the new information in e_t.

Dataset Construction

Our dataset:



- Composed of **knowledge facts** $\boldsymbol{c}$
- Fetch cited news from Common Crawl containing event info: $\mathrm{e} = (e_1, e_2, ..., e_t)$
- Sources: Wiki Current Event Portal and Common Crawl

Wikipedia version selection:

- **Original version** x_1 **:** Identified as the last edit **before event** $\boldsymbol{e_1}$.
- **Updated version** x_t **:** Identified as the first edit **after event** $\boldsymbol{e_t}$.

Dataset Construction

Total Instances: 5,695

Data Language: English

Distribution:

- Training: 80% (4,583 clusters, 276k articles)
- Validation: 10% (556 clusters, 43k articles)
- Test: 10% (556 clusters, 42k articles)

	Training	Validation	Test
# Clusters	4,583	556	556
# Articles (WCEP-100)	276k	43k	42k
Mean	60.3	76.8	74.9
Median	66	100	100
Period Begin	2016-8-25	2019-1-6	2019-5-8
Period End	2019-1-5	2019-5-7	2019-8-20

Methodology Overview

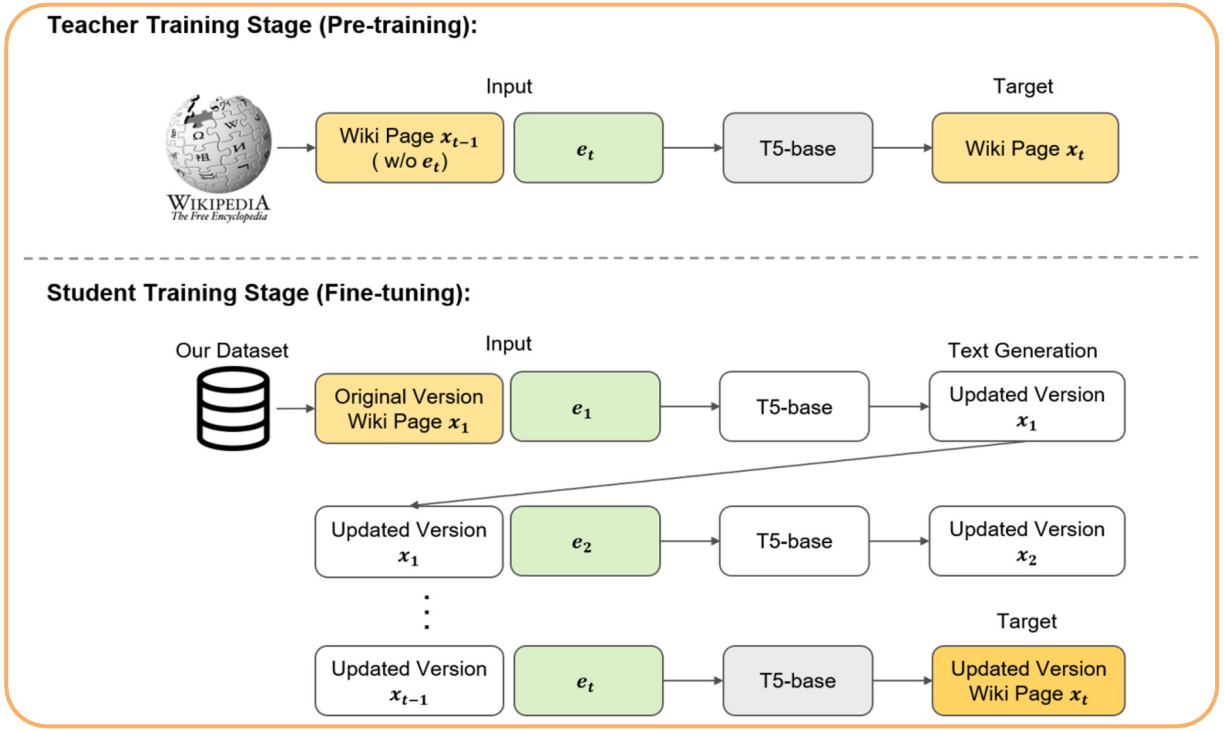
Hybrid Training Framework

- Teacher Training Stage (Pre-training)
- Student Training Stage (Fine-tuning)

Model

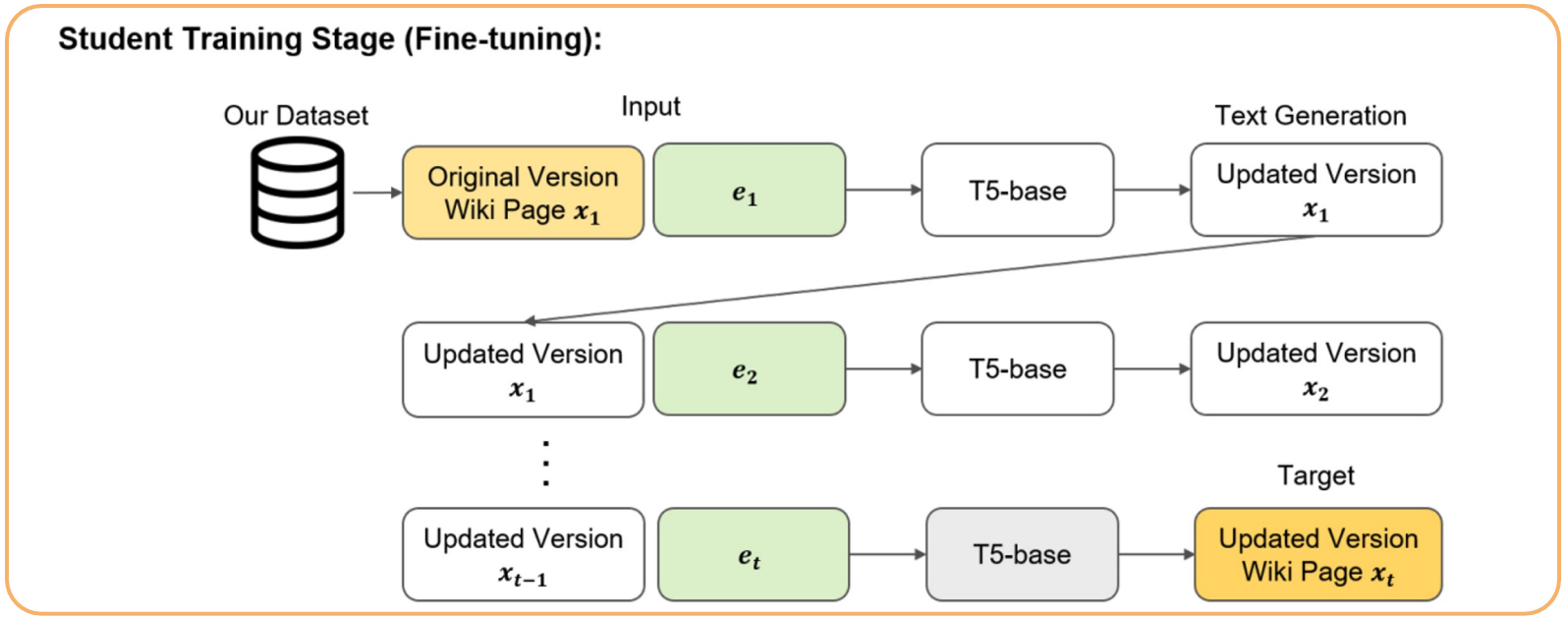
Text-To-Text Transfer Transformer (T5)

(L=12, H=768, A=12, 220M params)



Methodology - Student Training Stage

- Fine-tunes using our dataset.
- Iteratively updates text with new emerging information.



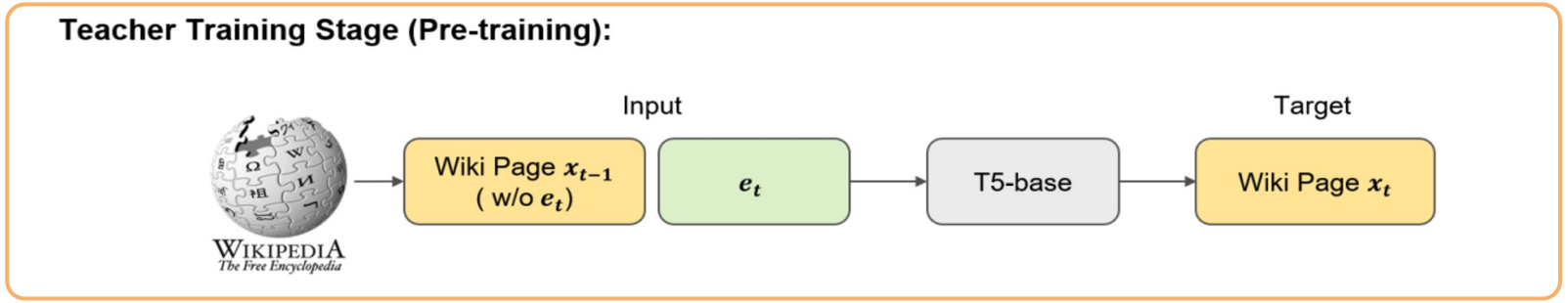
Methodology - Teacher Training Stage

1. Utilizes existing dataset **NetKu (labeled data)**.

2. Uses unlabeled data for **self-supervised training (unlabeled data)**.

Dataset	# Inst.	Avg. Sent.	Avg. Tokens
Labeled data (NetKu)	253	12.27	244.60
Unlabeled data (WCEP)	1,990	9.47	193.11
Total	2,243		

Methodology - Self-supervised Pre-training



Objective:

- Address data scarcity by treating textual knowledge updating as text reconstruction.

Procedure:

- Extract 15% sentences from a Wikipedia summary as new info x_t
- Remove duplicated sentences and form input using the format "**update:{** x_{t-1}**} event:{**e_t**}**"
- Use x_t as reference output
- Obtain the triplet of **(**x_{t-1}**,** e_t **,** x_t **)** as candidate instances

Methodology - Perturbation Strategies

① **Sentence-Shuffle (SS):** Shuffling order of original sentences in e_t .

② **Noise-Injection (NI):** Injecting 20% irrelevant noise into e_t .

③ **Noise-Generation (NG):** Generating 20% noise injection via **GENIUS** *[Guo et al., 2022]* into e_t .

Perturbation Strategy	Scenario in text reconstruction
Original (No Perturbation)	To update x_{t-1} (Wiki Content) with the event e_t (original news article)
Sentence-Shuffle (SS)	To update x_{t-1} with event e_t that is randomly shuffled
Noise-Injection (NI)	To update x_{t-1} with event e_t that is injected with noise from irrelevant Wiki content
Masked-Noise-Injection (MNI)	To update masked x_{t-1} with event e_t that is injected with noise from irrelevant Wiki content
Noise-Generation (NG)	To update x_{t-1} with event e_t that is augmented with noise generation
Masked-Noise-Generation (MNG)	To update masked x_{t-1} with event e_t that is augmented with noise generation

Methodology - Perturbation Strategies

④ **Masked-Noise-Generation (MNG):**

1. Based on **Noise-Generation**, with e_t unchanged, the noise is increased against x_{t-1}.
2. **Extracting keywords from x_{t-1}** via the YAKE algorithm used in the **GENIUS model**.
3. The remaining parts of the text are **replaced by mask tokens**, and the **masking ratio is based on 15%** of the original T5 pre-training.

⑤ **Masked-Noise-Injection (MNI):** Based on Noise-Injection, use the GENIUS model to mask x_{t-1}.

Perturbation Strategy	Scenario in text reconstruction
Original (No Perturbation)	To update x_{t-1} (Wiki Content) with the event e_t (original news article)
Sentence-Shuffle (SS)	To update x_{t-1} with event e_t that is randomly shuffled
Noise-Injection (NI)	To update x_{t-1} with event e_t that is injected with noise from irrelevant Wiki content
Masked-Noise-Injection (MNI)	To update masked x_{t-1} with event e_t that is injected with noise from irrelevant Wiki content
Noise-Generation (NG)	To update x_{t-1} with event e_t that is augmented with noise generation
Masked-Noise-Generation (MNG)	To update masked x_{t-1} with event e_t that is augmented with noise generation

Experiments - Pre-training Strategy Results

- Pre-training strategies enhance the generative model's performance in updating knowledge.
- Hybrid approaches outperform the traditional supervised-learning model.
- Best performance strategy: **Hybrid (Original / Noise-Injection + Masked-Noise-Generation)**
- Both noise types and original T5 masking ratio amplified knowledge update ability.

Method	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BERTScore
Student Mode	15.14	7.61	12.26	3.67	2.31	1.98	1.83	7.50	57.68
Hybrid (Original)	67.94	63.50	65.63	52.22	49.65	48.46	47.53	57.53	84.95
Hybrid (Original+SS)	74.02	70.46	72.05	63.22	61.05	59.82	58.85	66.93	87.85
Hybrid (Original+NI)	85.19	81.81	83.23	77.15	74.69	73.25	72.13	79.76	92.94
Hybrid (Original+MNI)	78.61	74.63	76.56	68.52	66.06	64.66	63.56	71.21	90.04
Hybrid (Original+NG)	83.32	79.86	81.39	74.97	72.50	71.09	69.95	77.71	92.03
Hybrid (Original+MNG)	84.45	80.88	82.38	76.64	74.07	72.73	71.70	79.31	92.62
Hybrid (Original/NI+MNG)	**86.44**	**83.02**	**84.46**	**79.63**	**77.08**	**75.65**	74.54	**82.16**	**93.49**

Experiments - Experiments with LLMs

Four state-of-the-art LLMs : GPT-3.5-turbo, LLaMA-13B, Vicuna-13B, Falcon-7B.

1. **LLMs Zero-Shot Experiment :**
   - **Enabling direct prediction by LLMs** for text sequence updates.
   - Using the Student Training Stage framework

2. **LLMs Fine-Tuning Experiment:**
   - Using the best self-supervised combination: **Hybrid (Original / Noise-Injection + Masked-Noise-Generation).**
   - Four LLMs fine-tuned using the **LoRA method**.

Experiments - LLMs Experimental Results

- Our model outperforms GPT-3.5-turbo, LLaMA-13B, Vicuna-13B, Falcon-7B in all metrics.
- Zero-shot GPT-3.5-turbo commendable but not as efficient as our model.
- Post-LoRA fine-tuning:
  - Performance increase across all LLMs.
  - Vicuna-13B's efficiency close to GPT-3.5-turbo.
- Highlights efficacy of our training strategy.

Method	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BERTScore
Student Mode	15.14	7.61	12.26	3.67	2.31	1.98	1.83	7.50	57.68
Hybrid (Original)	67.94	63.50	65.63	52.22	49.65	48.46	47.53	57.53	84.95
Hybrid (Original/NI+MNG)	**86.44**	**83.02**	**84.46**	**79.63**	**77.08**	**75.65**	**74.54**	**82.16**	**93.49**
GPT-3.5-turbo (Zero-shot)	49.35	34.98	40.41	39.79	31.71	30.07	29.59	41.00	72.75
Falcon-7B (Zero-shot)	15.96	2.08	10.14	8.16	0.99	0.13	0.02	8.36	42.42
LLaMA-13B (Zero-shot)	16.60	9.28	13.40	9.53	7.57	7.31	7.26	10.77	47.25
Vicuna-13B (Zero-shot)	43.94	28.68	35.67	36.18	26.75	25.17	24.89	35.97	68.16
Falcon-7B (Fine-Tuned)	21.20	4.66	12.32	10.79	2.19	0.53	0.18	11.13	55.38
LLaMA-13B (Fine-Tuned)	32.53	22.55	27.66	24.01	17.72	16.32	15.66	30.19	60.00
Vicuna-13B (Fine-Tuned)	48.00	36.25	41.76	42.00	34.28	33.22	33.04	42.55	68.98

Experiments - Example of Comparison

Comparison of model outputs:

- LLMs often produce lengthy, redundant outputs.
- Our model's outputs are precise and topical.

Our model	In August 2017, reports surfaced that American and Canadian diplomatic personnel in Cuba had suffered a variety of health problems, dating back to late 2016, and accusations were made that these were a result of an attack by someone using unspecified technology, possibly sonic in nature. The media has taken to calling these reported incidents ""Havana Syndrome"". The Cuban government was at first exonerated by the U.S. government, but then in a reversal, Cuba was accused of perpetrating unspecified attacks causing these symptoms. The U.S. reduced staff at their embassy to a minimum, and U.S. President Trump declared in October 2017 that he believed Cuba was responsible for the attacks, but offered no evidence for his claim. Others expressed doubts, including scientific skeptics such as Brian Dunning, at least one U.S. senator, the director of the Cuban Neuroscience Center, and mass psychogenic illness expert Robert Bartholomew. In April 2018, U.S. diplomats in China began to report problems similar to those reported in Cuba.
GPT-3.5-turbo	Beginning in late 2016, a mysterious illness befell U.S. diplomatic personnel and family members in Havana, Cuba. The affected individuals reported severe physical symptoms, including vertigo, headaches, ear pain, cognitive difficulties, and hearing loss, after reportedly hearing mysterious noises in their hotels or homes. In August 2017, reports surfaced that American and Canadian diplomatic personnel in Cuba had suffered from these symptoms, and the media coined the term "Havana Syndrome" to refer to these incidents. The Cuban government was initially exonerated by the U.S. government, but in a reversal, Cuba was later accused of perpetrating unspecified attacks causing these symptoms. The U.S. reduced staff at their embassy to a minimum, and in October 2017, U.S. President Trump claimed Cuba was responsible for the attacks but offered no evidence for his claim. Others, including scientific skeptics such as Brian Dunning, at least one U.S. senator, the director of the Cuban Neuroscience Center, and mass psychogenic illness expert Robert Bartholomew, expressed doubts about the claims. In April 2018, U.S. diplomats in China began to report problems similar to those reported in Cuba.In subsequent years, new research was conducted into the cause of the Havana Syndrome. In June 2019, a study by the National Academy of Sciences found that the symptoms experienced by U.S. diplomats were consistent with "directed, pulsed radiofrequency energy," a type of microwave radiation. In December 2020, a report by the National Academy of Medicine stated that "directed, pulsed radiofrequency energy" was the most likely cause of the illness.However, in May 2021, the theory of a directed energy attack was challenged by a report in The New Yorker, which found that the sound heard by the U.S. embassy workers in Cuba matched the chirp of a particular species of cricket ...

Conclusion

- This study tackles real-time **sequential knowledge update** by extracting information from constantly updating news sources.

- It proposes a **hybrid learning architecture** and **self-supervised training strategy** for language models **to consolidate knowledge like humans**.

- We have created a new dataset that **provides additional research resources** in the field.

Thank you!