Dynamic Model Switching to Mitigate Outdated Knowledge in Large Language Models

Ramakrishna Pinninti[1], **Sabyasachi Kamila**[2],
Ayan Mazumder[3], **Mohammed Hasanuzzaman**[1]
[1]ADAPT Centre, Computer Science Department, Munster Technological University, Cork, Ireland
[2]Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India
[3]IBM, USA
ramakrishna.pinninti@adaptcentre.ie

Abstract

Generating timely and accurate content is a significant challenge for Large Language Models (LLMs). Obsolete information reduces their reliability and user trust. To overcome the limitations of single models in adapting to evolving information, we propose a dynamic switching model. A multitask trained switch model objective, adaptively picks between a large model that does not have recent information and a smaller model fine-tuned on recent information using contextual and temporal indicators. This method incorporates semantic update detection and temporal switching, which predicts text obsolescence through aggregation of reward signals. For evaluation, we curated the Temporally-aware Dynamic Dataset (TaDD) on Wikipedia and Guardian articles, which are frequently updated. Our framework achieves a balanced precision-recall trade-off on five datasets without continuous retraining, which shows that the model is efficient and adaptable compared to static pretrained models.The code and dataset are publicly available in our GitHub repository. our GitHub repository.

Keywords: Text Obsoleteness, LLM, Reinforcement learning, MultiTask learning

1. Introduction

Maintaining the timeliness and accuracy of content in large-scale knowledge repositories is crucial outdated information can reduce both trust and usability. LLMs, typically trained or fine-tuned on static datasets, often struggle to keep up with continuously evolving knowledge. Detecting obsolescence and predicting content validity have, as a result, become important research challenges (Jatowt et al., 2013; Wenzel and Jatowt, 2024). Traditional text generation approaches face difficulties in balancing broad knowledge coverage with the ability to adapt to newly emerging information (Hosokawa et al., 2024). For example, a large but static model (Big_Model), trained on extensive datasets, usually possesses wide general knowledge but lacks awareness of recent developments (Lee et al., 2025). In contrast, a smaller and frequently updated model (Small_Model) can better capture recent information but typically suffers from limited coverage and contextual understanding. Fine-tuning a single model is a commonly used strategy, but it frequently fails to remain consistently up-to-date. Ensemble approaches, such as averaging outputs from multiple models, can introduce inconsistencies when the models rely on conflicting temporal contexts (Zhang et al., 2025). To address these challenges, we propose a dynamic model-switching framework that adaptively selects between Big_Model (representing historical knowledge) and Small_Model (representing recent knowledge) based on the input context. Unlike static or ensemble-based methods (Ranaut et al., 2025), our dynamic switch model routes each input to the most temporally appropriate model in real time. This way, the generated content stays factually accurate and temporally consistent (Wenzel and Jatowt, 2023). On top of that, a temporal validity evaluation mechanism tightens the model selection process by weighing contextual relevance rewards (r_{context}) against an evolving threshold (τ), yielding more adaptive and context-aware outputs. Our framework is specifically designed for detecting content spans with a high likelihood of obsolescence, supporting practical applications such as highlighting potentially outdated content in web browsers, assisting fact-checking pipelines, and improving content selection for training LLMs. Unlike existing datasets that focus solely on semantic updates, we construct a Temporally–aware Dynamic Dataset integrating model outputs, temporal validity labels, and decision probabilities, annotated by human with assistance from GPT-4.1 Mini (OpenAI et al., 2024). By dynamically leveraging both models through a switch, our approach ensures contextually coherent, and temporally accurate text generation, providing a robust benchmark for advancing temporally–aware language models.

2. Related Work

Temporal Reasoning in LLMs. Large Language Models (LLMs) lack intrinsic temporal awareness, making it challenging to generate factually accurate and up-to-date content. Sojitra et al. (2024) ex-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2490–2500
11–16 May 2026.

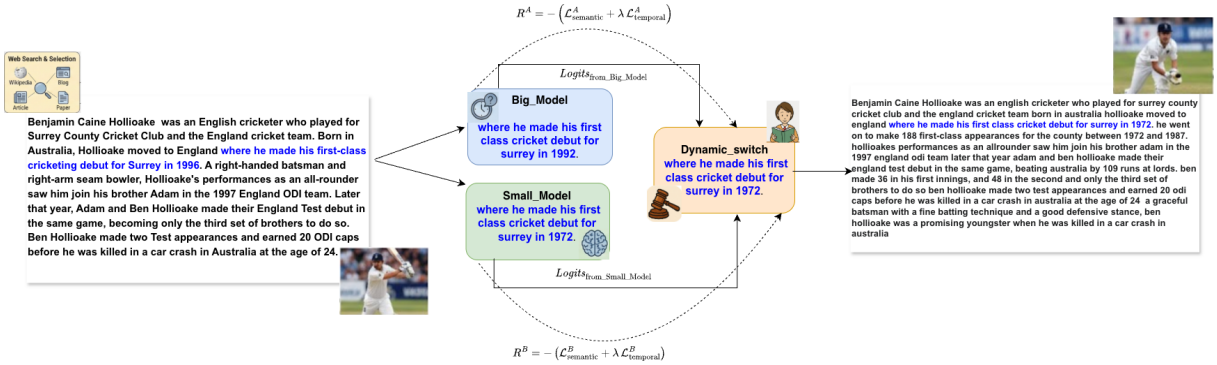

Figure 1: A schematic of our proposed dynamic model-switching framework. Given an input x_1, the framework evaluates two temporally specialized models: M_1 (Big_Model), trained on historical data, and M_2 (Small_Model), fine-tuned on recent data. The dynamic switch model selects the most temporally relevant model M^* based on contextual relevance reward r_{context} and a dynamically updated temporal validity threshold τ. The reward functions R^A and R^B integrate semantic correctness $\mathcal{L}_{\text{semantic}}$ and temporal validity $\mathcal{L}_{\text{temporal}}$, ensuring factually accurate and temporally coherent text generation. The switch model continuously optimizes decisions based on past performance, selecting the model that maximizes overall reward.

plored timestamped datasets and structured knowledge to enhance temporal summarization, but adapting to evolving real-world contexts remains difficult. Wallat et al. (2024) showed that LLMs struggle with time-sensitive question answering, especially with event sequencing, real-time updates, and maintaining temporal consistency. Zhou et al. (2019) introduced MCTACO, a dataset for evaluating temporal commonsense reasoning across event duration, ordering, frequency, and stationarity, and found significant gaps compared to human performance. Yang et al. (2017) created a taxonomy for classifying semantic edit intentions in Wikipedia revisions, showing that copy editing and wikification improve retention, while elaboration and verification support article quality. Jain et al. (2023) benchmarked eight LLMs across six datasets, finding that while models handle some tasks, they still struggle with event ordering, multiple-event reasoning, and precise time prediction.

Temporal Robustness. Wallat et al. (2025) analysed the temporal robustness of LLMs and demonstrated that models often struggle to correctly interpret time-sensitive queries. They introduced eight diagnostic tests, including temporal reversal, year shifting, and time relativisation, to analyse how temporal transformations affect model performance. The authors further show that simple question reformulations can improve temporal QA performance by up to 55%. Their findings show that model performance drops when the temporal expressions are modified. This is memorising rather than temporal reasoning.

Decoding Strategies. Li et al. (2023) propose contrastive decoding, a decoding strategy aimed at improving the quality of text generated by large language models. Standard methods, including greedy search, top-k sampling, and nucleus sampling, frequently generate incoherent and repetitive text. To address these limitations, the authors proposed a framework that contrasts predictions from the expert model with the help of an amateur model, prioritising tokens under the guidance of the expert model but less likely under the weaker model.

Dynamic Adaptation. Wang et al. (2025) introduced LLM-DA, a dynamic adaptation method for temporal knowledge graph reasoning (TKGR). It extracts logical temporal information from historical data and dynamically updates it, thereby eliminating the need for expensive fine-tuning. This approach has provided a balance between explainability and adaptability, with strong results across a wide range of datasets. Piryani et al. (2025) provides a comprehensive survey of temporal information retrieval and question answering, emphasizing challenges in temporal intent, event ordering, and fact evolution reasoning. They compared traditional neutral methods and transformer-based LLMs, describing their robustness, and timeliness.

Sentence Validity Prediction. Almquist and Jatowt (2019) studied sentence validity prediction, aiming to estimate how long textual statements remain valid. Their work used linguistic and temporal cues to determine content expiry periods, supporting time-sensitive information retrieval and updating tasks.

Cao and Wang (2022) analysed time-based prompting of text generation by using document timestamps as textual or continuous representations. They demonstrated that using large-scale, chronologically ordered datasets, including TEMPWIKIBIO, enhanced temporal coherence and factual alignment through timestamp conditioning. Nevertheless, unlike these single-model prompting

models, our framework dynamically directs inputs to temporally specialized generators.

Model Editing. Model editing methods aim to efficiently update LLM knowledge. Mitchell et al. (2022) introduced SERAC, a memory-based framework for integrating new facts at inference. More recent work, *Memory-Based Model Editing at Scale*, extends this paradigm by enabling scalable, localized edits without retraining. While effective for factual correction, editing methods are less suited for balancing temporal validity across evolving contexts.

Our work bridges this gap by introducing a dynamic model-switching framework that learns to select between a temporally specified model and frequently updates it by aggregating reward histories. This effectively captures text obsolescence by comparing present and past reward signals (see Table 1). Using a curated Temporally-aware Dynamic Dataset (TaDD), our method enables proactive handling of content obsolescence without the need for continual retraining.

Dataset	Temporal Update	Semantic Update	Total Samples
Edit-Intentions (Spangher et al., 2024)	1,264	3,290	4,554
Wiki-TIDE (Borkakoty et al., 2023)	1,590	548	2,138
One Document, Many Revisions (Rajagopal et al., 2022)	312	0	312
TaDD (Ours)	**1,081**	**1,419**	**2,500**
Total	**4,247**	**5,257**	**9,504**

Table 1: Statistics of Temporal and Semantic updates across benchmark datasets and our TaDD dataset.

3. Method

The dynamic model-switching framework is designed to adaptively select the most suitable model from a set of models, $\{M_1, M_2\}$, where each model is fine-tuned on datasets from different temporal contexts. For example, M_1 specializes in generating outputs based on outdated or historical data (typically being a larger model due to having more available data for training)[1]. In contrast, M_2 is optimized for handling recent and up-to-date information[2].Given an input x_1, the objective is to dynamically identify and select the optimal model M^* that maximizes both contextual relevance and temporal validity of the generated output.

Our problem is formulated as minimizing a total loss function, which integrates two key components: $\mathcal{L}_{\text{switch}}$ and $\lambda\mathcal{L}_{\text{temporal}}$. The $\mathcal{L}_{\text{switch}}$ ensures that the framework learns to select the best model M^* for each input x_1, leveraging the contextual relevance of the generated output. The $\mathcal{L}_{\text{temporal}}$ evaluates whether the output generated by the selected model aligns with the temporal requirements of the input. λ is a scalar weighting factor that controls the contribution of the temporal validity loss to the total loss.

The objective can be written mathematically as:

$$\mathcal{L}_{\text{total}} = \underbrace{-\sum_{i=1}^{2} y_i \log\big(\pi(i \mid z_{\text{concat}})\big)}_{\text{Model Selection}} + \underbrace{\lambda\left[-\big(v \log(\hat{v}) + (1-v)\log(1-\hat{v})\big)\right]}_{\text{Temporal Validity}} \quad (1)$$

Model Selection

The proposed dynamic switching mechanism selects the most suitable model for each input. The model selection loss ensures that the switch model learns to correctly predict the optimal model M^*, which maximizes the reward function $R(M_i, x)$ for a given input x.

$$R(M_i, x) = 1 - \mathcal{L}_{\text{total}}(M_i, x) \quad (2)$$

This equation defines the reward $R(M_i, x)$ for a model M_i on input x as the complement of the total loss $\mathcal{L}_{\text{total}}(M_i, x)$.

A lower total loss corresponds to a higher reward, encouraging the model to produce outputs that are both semantically accurate and temporally valid.

$$M^* = \underset{M_i \in \{M_1, M_2\}}{\arg\max} \; R(M_i, x) \quad (3)$$

$$\mathcal{L}_{\text{switch}} = -\sum_{i=1}^{2} y_i \log\big(\pi(i \mid z_{\text{concat}})\big) \quad (4)$$

y_i: A one-hot encoded label indicating the optimal model, determined by the reward (e.g., if M_1 is better, $y = [1, 0]$). $\pi(i \mid z_{\text{concat}})$: The probability of selecting model M_i, computed using the softmax function:

$$\pi(i \mid z_{\text{concat}}) = \frac{\exp(\text{logit}_i)}{\sum_{j=1}^{2} \exp(\text{logit}_j)} \quad (5)$$

Semantic Loss ($\mathcal{L}_{\text{semantic}}$). In addition to the switching objective, semantic correctness of the generated sequence is enforced using a standard cross-entropy loss between the predicted tokens and the reference output:

$$\mathcal{L}_{\text{semantic}} = -\sum_{t=1}^{T} \log p(\hat{y}_t \mid \hat{y}_{<t}, x), \quad (6)$$

where $p(\cdot)$ denotes the probability distribution over tokens produced by the selected model M^*, $\hat{y}_t$ represents the predicted token at time step t, and y_t denotes the corresponding ground-truth token in the reference sequence. This term ensures that the model not only selects the correct temporal expert but also generates factually and linguistically accurate text consistent with the reference output.

[1] https://huggingface.co/microsoft/Phi-3-medium-128k-instruct

[2] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

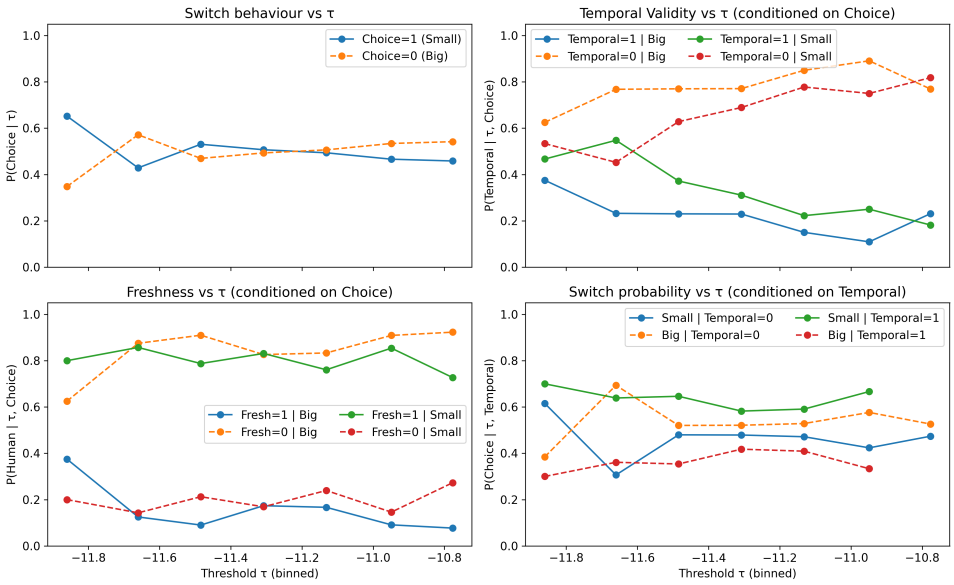

Figure 2: Sensitivity analysis of our dynamic switching framework illustrating the interaction between the adaptive threshold τ, switching decisions (Big_Model vs Small_Model), temporal validity, and freshness. Switching probabilities vary smoothly across τ bins, reflecting controlled threshold sensitivity. Model selection is not deterministic; instead of strictly mapping outdated and updated content to fixed models, the switch adapts decisions based on contextual reward signals.

Algorithm 1 Dynamic Model Switching Framework

Require: Input x, Models M_1, M_2, Threshold τ, Switch model π
Ensure: Selected output y

Compute $z_{\text{concat}} = \text{concat}(M_1(x), M_2(x))$ ▷ concatenation of logits from M_1 and M_2, which serves as input to the switch model.
Compute switch probabilities: $\pi(i \mid z_{\text{concat}})$
Select model: $M^* = \arg\max_i \pi(i \mid z_{\text{concat}})$
Generate output: $y = M^*(x)$
Compute contextual reward r_{context}
if $r_{\text{context}} \geq \tau$ **then**
    Mark output as **Temporally Valid**
else
    Mark output as **Temporally Invalid**
end if
Update τ: $\tau \leftarrow \alpha \cdot \tau + (1 - \alpha) \cdot r_{\text{context}}$

Temporal Validity Assessment

Temporal validity is calculated based on the reward history to ensure that the generated output aligns with the temporal context of the input. It involves dynamically updating a temporal validity threshold (τ) that evolves over time, utilizing past rewards. The threshold serves as a criterion for determining whether the generated output is temporally valid or invalid, allowing our dynamic switch to adjust its temporal alignment decisions in response to evolving model performance.

$$\mathcal{L}_{\text{temporal}} = -\big(v \cdot \log(\hat{v}) + (1 - v) \cdot \log(1 - \hat{v})\big) \quad (7)$$

Temporal validity is evaluated by comparing the contextual relevance reward (r_{context}) with the dynamically updated temporal validity threshold (τ). This evaluation results in a binary decision, classifying outputs as either temporally valid ($v = 1$, if $r_{\text{context}} \geq \tau$) or temporally invalid ($v = 0$, if $r_{\text{context}} < \tau$), $\hat{v}$: The predicted probability of temporal validity is computed as $\hat{v} = \sigma(\text{temporal\_logits})$, where σ is the sigmoid function.

The dynamically updated threshold ensures that temporal validity decisions adapt to the reward history, reflecting the models' recent performance. This mechanism keeps the system robust, context-aware, and aligned with the temporal requirements of the input.

The temporal validity threshold (τ) is updated at each time step using a weighted running average of the most recent contextual reward (r_t) and the previous threshold (τ_t):

$$\tau_{t+1} = \alpha \cdot \tau_t + (1 - \alpha) \cdot r_t \quad (8)$$

Here τ_{t+1} is the updated threshold at time step $t + 1$, τ_t is the threshold at the current time step t, r_t is the most recent contextual relevance reward, and α is the smoothing factor ($0 < \alpha < 1$) that controls the balance between past thresholds and recent rewards. As shown in Figure 2, switching probabilities vary smoothly across different τ bins, indicating that the model selection mechanism is reward-sensitive and dynamically adaptive rather than governed by a static mapping between outdated and updated content.

S. No.	Wiki-ref	Old Content	New Content	Human Tag
1	**Ben Hollioake**	Hollioake Benjamin Caine was an English cricketer who played for Surrey County Cricket Club and the England cricket team. Born in Australia, Hollioake moved to England where he made his first-class cricket debut for Surrey in 1996. A right-handed batsman and right-arm seam bowler.	Hollioake Benjamin Caine was an English cricketer who played for the Surrey County Cricket Club and the England cricket team. Born in Australia, Hollioake moved to England where he made his first-class cricket debut for Surrey in 1972. He went on to make 188 first-class appearances for the county between 1972 and 1987.	1
2	**Nichols Algebra**	This root system is the smallest member of an infinite series. The images are from ref name=CL15, where this example is alsodiscussed in detail.	This root system represents the smallest element of an infinite series. The images originate from ref name=CL15, where this example is also explained thoroughly.	0
3	**sports**	In 1994 hollioake made his first class debut for surrey in 1996 taking four 4–74 against yorkshire at acklam park middlesbrough and was awarded the nbc denis compton award.	In 1994 hollioake made his first class debut for surrey in 1996 taking four 4–74 against yorkshire at acklam park middlesbrough and was awarded man of the match. he was awarded a county cap in 1999. He played for Surrey in 2001 and took 20 wickets in the season.	1

Table 2: Examples from our Temporally-aware Dynamic dataset. Updated content is highlighted in blue.

4. Datasets

We used several existing datasets and created a Combined Dataset that includes 8,085 sentences from Edit-Intentions (Spangher et al., 2024), Wiki-TIDE (Borkakoty and Espinosa-Anke, 2023), and One Document, Many Revisions (Rajagopal et al., 2022), serving as the foundation for our dynamic dataset augmentation pipeline, and created a Temporally-aware Dynamic Dataset (TaDD). Although these datasets are valuable for analyzing semantic updates, they lack explicit temporal context awareness, dynamic decision evaluation, and alignment with real-world temporal challenges. Our TaDD dataset bridges these gaps by leveraging a dynamic switching mechanism that selects between two temporally fine-tuned models (M_1 for historical data and M_2 for recent data) to produce outputs optimized for both semantic relevance and temporal validity.

Our TaDD dataset is curated using Wikipedia pages and Guardian News Articles[3] spanning 2016–2024, covering diverse domains such as sports, politics, health, and technology, ensuring domain diversity and temporal breadth. Each example includes prompts and corresponding ground truths, outputs generated by M_1, M_2, and the chosen model (M^*), temporal validity labels (v), dynamic thresholds (τ), and model selection probabilities ($\pi(i \mid z_{\text{concat}})$). The Big_Model ($M_1$) and Small_Model ($M_2$) remain fully frozen throughout all experiments and are not trained or fine-tuned on TaDD instances. Their generated outputs are stored solely as auxiliary metadata for analyzing routing behavior, and training and evaluation splits are strictly separated to ensure that performance reflects the effectiveness of the dynamic switching mechanism rather than adaptation of the underlying generators.

[3] https://www.kaggle.com/datasets/adityakharosekar2/guardian-news-articles

The existing datasets provided a starting point for semantic annotations, which we validated and refined through a two-step annotation process involving GPT-4.1 Mini followed by human validation, focusing on significant factual changes such as numerical updates, dates, and status changes. Temporal alignment is evaluated by comparing the reward of the selected model with a dynamically updated threshold, ensuring that the generated outputs reflect evolving temporal relevance. By integrating semantic and temporal evaluations with dynamic decision metadata, our dataset provides a robust benchmark for advancing dynamic, temporally-aware text generation systems.

To ensure the accuracy of our dynamic switching framework, we randomly selected 1,000 test samples from the TaDD dataset. These samples were used for human evaluation with GPT-4.1 Mini, which was used solely to assist annotation (not as a validator). For human evaluation, the test samples were assigned to two linguistically proficient annotators, who classified each pair of outdated and updated sentences as either a temporal update or a semantic update. A pair was labeled as a temporal update if it contained substantial factual modifications such as changes in numerical values, dates, scores, episode counts, or status updates. In contrast, minor editorial changes, including rewording, punctuation, or stylistic edits, were categorized as semantic updates. Annotators followed a structured prompt instructing them to assess whether the updated text introduced new factual information and to return a binary classification of “Yes” or “No”. The human annotations served as the gold standard; GPT-4.1 Mini outputs were used only as drafting hints and to assess performance and scalability, following established practices from LLM reliability studies. For annotation assistance, we prompted GPT-4.1 Mini to suggest whether a given sentence would likely undergo a temporal update in the near future. The

model was assigned a similar classification suggestion task, indicating whether the content was dynamic and subject to future change based on its factual nature. The prompt instructed GPT-4.1 Mini to return “Yes” if it contained dynamic information (e.g., dates, scores, statuses), or “No” if the content was static and unlikely to change. Final labels were assigned by human annotators; model suggestions were advisory and did not determine ground truth. By combining expert human annotation with GPT-4.1 Mini–assisted drafting, we ensured a rigorous and consistent process for identifying temporal updates.

5. Results and Analysis

Table 10 presents the performance metrics for the Temporal Validity Assessment on the Combined Dataset, evaluating the models based on accuracy, precision, recall, and F1-score. The evaluation is carried out using the Combined dataset, including Edit-Intentions, Wiki-TIDE, and One Document, Many Revisions. In this dataset, entries labeled as "Fact Updates" were classified as Label=1 (temporal update), while all others were marked as Label=0 (semantic update). Wiki-TIDE entries with minor and fundamental changes were initially annotated using GPT-4.1 Mini, followed by human validation. To ensure robust training, we used the Combined Dataset (mentioned in Section 4) and allocated 80% of instances for training and 20%.

Model	Accuracy (Δ)	Precision (Δ)	Recall (Δ)	F1-Score (Δ)	Active Params (Train / Infer)
GPT-4.1 Mini (Few-shot)	0.78 (↑20%)	0.65 (↓2%)	0.72 (↑11%)	0.68 (↑4%)	0 / 14B
GPT-4.1 Mini (Zero-shot)	0.75 (↑15%)	0.60 (↓9%)	0.70 (↑8%)	0.65 (↓1%)	0 / 14B
LLaMA-3 (8B) Few-shot	0.65 (=0%)	0.70 (↑6%)	0.93 (↑43%)	0.80 (↑22%)	0 / 8B
LLaMA-3 (8B) Zero-shot	0.61 (↓6%)	0.57 (↓14%)	0.77 (↑18%)	0.66 (↑1%)	0 / 8B
LLaMA-2-7B Zero-shot	0.58 (↓11%)	0.57 (↓14%)	0.59 (↓9%)	0.58 (↓11%)	0 / 7B
Mistral-7B Zero-shot	0.48 (↓26%)	0.47 (↓29%)	0.60 (↓8%)	0.53 (↓19%)	0 / 7B
LLaMA-2-7B Fine-Tuned	0.55 (↓15%)	0.50 (↓24%)	0.59 (↓9%)	0.54 (↓18%)	7B / 7B
Mistral-7B Fine-Tuned	0.75 (↑15%)	0.72 (↑9%)	0.78 (↑20%)	0.75 (↑15%)	7B / 7B
GPT-2 Fine-Tuned	0.72 (↑11%)	0.75 (↑14%)	0.70 (↑8%)	0.72 (↑10%)	124M / 124M
BART Fine-Tuned	0.69 (↑6%)	0.62 (↓6%)	0.64 (↓2%)	0.63 (↓4%)	139M / 139M
Our Model (Dynamic-Switch)	**0.65** (=0%)	**0.66** (=0%)	**0.65** (=0%)	**0.65** (=0%)	**120M / (14B + 3.8B)**
Phi-3 Mini Fine-Tuned	0.70 (↑8%)	0.62 (↓6%)	0.74 (↑14%)	0.67 (↑2%)	3.8B / 3.8B
Phi-3 Mini Zero-shot	0.55 (↓15%)	0.50 (↓24%)	0.79 (↑22%)	0.61 (↓6%)	0 / 3.8B
Phi-3 Medium (14B) Few-shot	0.67 (↑3%)	0.65 (↓2%)	0.73 (↑12%)	0.69 (↑5%)	0 / 14B
Phi-3 Medium (14B) Zero-shot	0.65 (=0%)	0.60 (↓9%)	0.72 (↑11%)	0.65 (=0%)	0 / 14B

Table 3: Performance comparison on the Combined Temporal Validity dataset. Colored arrows (↑, ↓, =) indicate percentage change relative to our Dynamic-Switch model. Active Params (Train / Infer) shows parameters used for training and inference, respectively.

Our dynamic framework achieves an accuracy of 0.65 on the Combined Temporal Validity dataset. While fine-tuned static models, such as LLaMA-2-7B Fine-Tuned (0.55) and Mistral-7B Fine-Tuned (0.75), demonstrate competitive accuracy, zero–shot variants, including LLaMA-2-7B Zero–shot (0.58) and Phi-3 Mini Zero–shot (0.55), continue to suffer from poor precision–recall tradeoffs. For example, although Phi-3 Mini Zero–shot shows high recall (0.79), its low precision (0.50) results in over-detection of temporal changes. Similarly, LLaMA-3 (8B) Zero–shot achieves a strong recall (0.77), but has lower accuracy (0.61) and precision (0.57). In contrast, our dynamic selection framework achieves a precision of 0.66 and a recall of 0.65 without requiring retraining, resulting in fewer outdated responses and improved generalizability. Static fine-tuned models may benefit from retraining, but they lack real-time adaptability. Even high-performing models like GPT-2 Fine-Tuned (accuracy 0.72) and BART Fine-Tuned (0.69) cannot match the balanced performance of our dynamic switch, especially in evolving-knowledge scenarios.

We also observe that GPT-4.1 Mini (Few–shot) achieves the highest accuracy (0.78) with strong *F1* (0.68), and LLaMA-3 (8B) Few–shot excels in recall (0.93) and *F1* (0.80), outperforming static models on isolated metrics. However, few–shot

Model	Accuracy (Δ)	Precision (Δ)	Recall (Δ)	F1-Score (Δ)
GPT-4.1 Mini (Zero-shot)	0.75 (↑7%)	0.60 (↓3%)	0.70 (↓5%)	0.65 (↓4%)
GPT-4.1 Mini (Few-shot)	0.78 (↑11%)	0.65 (↑5%)	0.72 (↓3%)	0.68 (↑1%)
LLaMA-3 (8B) Zero-shot	0.61 (↓13%)	0.57 (↓8%)	0.77 (↑4%)	0.66 (↓3%)
LLaMA-3 (8B) Few-shot	0.65 (↓7%)	0.70 (↑13%)	0.93 (↑26%)	0.80 (↑19%)
Phi-3 Medium (14B) Zero-shot	0.65 (↓7%)	0.60 (↓3%)	0.72 (↓3%)	0.65 (↓4%)
Phi-3 Medium (14B) Few-shot	0.67 (↓4%)	0.61 (↓2%)	0.73 (↓1%)	0.67 (=0%)
LLaMA-2-7B Fine-Tuned	0.57 (↓13%)	0.58 (↓6%)	0.60 (↓19%)	0.59 (↓13%)
Mistral-7B Fine-Tuned	0.57 (↓13%)	0.58 (↓6%)	0.59 (↓20%)	0.58 (↓14%)
GPT-2 Fine-Tuned	0.54 (↓23%)	0.54 (↓13%)	0.65 (↓12%)	0.59 (↓13%)
BART Fine-Tuned	0.53 (↓24%)	0.54 (↓13%)	0.65 (↓12%)	0.59 (↓13%)
Our Model (Dynamic-Switch)	**0.70** (=0%)	**0.62** (=0%)	**0.74** (=0%)	**0.67** (=0%)
Phi-3 Mini Fine-Tuned	0.50 (↓29%)	0.49 (↓21%)	0.71 (↓4%)	0.58 (↓14%)

Table 4: Comparative performance on the Temporal Validity Assessment task using the TaDD dataset. Colored arrows (↑, ↓, =) indicate change relative to our Dynamic-Switch model.

Model	BLEU ↑	ROUGE-L ↑	Perplexity ↓
Static-Base Model (Big_Model)	23.7	35.2	20.5
Static-Update Model (Small_Model)	**28.1**	**34.6**	**19.3**
SERAC (Mitchell et al., 2022)	22.4	32.8	21.8

Table 5: Comparison of Big_Model (static base), Small_Model (updated static), and SERAC (Mitchell et al., 2022) across BLEU, ROUGE-L, and Perplexity metrics. Higher BLEU/ROUGE and lower Perplexity indicate better overall generation quality. Results show that the updated Small_Model achieves stronger fluency and temporal alignment, outperforming both the static and memory-based baselines.

Model	Accuracy ↑	Precision ↑	Recall ↑	F1-Score ↑
Static-Base Model (Big_Model)*	0.58	0.56	0.60	0.58
Static-Update Model (Small_Model)*	0.64	0.53	0.58	0.55
SERAC (Mitchell et al., 2022)	0.61	0.59	0.63	0.61
Dynamic-Switch Model	**0.65**	**0.66**	**0.65**	**0.65**

Table 6: Performance comparison of Static-Base, Static-Update, and Dynamic-Switch models using Accuracy, Precision, Recall, and F1-Score. Static models generated a single output per revision pair and were compared against the ground-truth freshness label (1 = updated, 0 = unchanged). Accuracy indicates overall correctness, while precision and recall capture the balance between correct and over-detected updates. *Static models were evaluated without reward histories or dynamic thresholds.

inference often requires more prompt engineering and computational resources. Our method maintains a favorable trade-off across metrics with minimal overhead, offering a lightweight, adaptable, and modular alternative. As shown in Table 10, our approach outperforms several baselines and closely rivals high-resource models, making it a scalable solution for freshness-critical tasks. Unlike fine-tuning or model editing, which risk *catastrophic forgetting* and high retraining costs, the dynamic switching model keeps generators frozen and trains only a lightweight 120M-parameter switch. This enables efficient temporal adaptation without performance drift, achieving a balanced precision–recall (0.66/0.65) at a far lower computational cost than few–shot or high-resource systems. Our model achieves a competitive *F1* (0.65) while avoiding the high computational cost of retraining, offering an efficient alternative to top-performing but resource-intensive few-shot models, such as GPT-4.1 Mini.

As shown in Table 4, our dynamic model selection framework achieves accuracy 0.70 on the TaDD dataset, outperforming all evaluated baselines. The evaluation is conducted using our TaDD Dataset, where '0' indicates no factual change and '1' represents an updated or revised fact. Compared to static fine-tuned models such as LLaMA-2-7B Fine-Tuned (0.57), Mistral-7B Fine-Tuned (0.57), and GPT-2 Fine-Tuned (0.54), our approach delivers consistent improvements across all metrics without requiring retraining. Our model maintains a balanced trade-off between precision 0.62 and recall 0.74, outperforming other models that tend to skew toward either high recall or high precision. For instance, GPT-2 Fine-Tuned achieves decent recall (0.65) but low precision (0.54), while Phi-3 Mini Fine-Tuned shows relatively high recall (0.71) but poor precision (0.49), leading to an over-detection of temporal updates.

Even among advanced models, trade-offs persist. GPT-4.1 Mini (Few–shot) achieves the highest accuracy (0.78) but slightly underperforms in recall (0.72) compared to our model. LLaMA-3 (8B) Few–shot yields extremely high recall (0.93) and strong *F1* (0.80), but lower accuracy (0.65), suggesting aggressive update detection. Meanwhile, LLaMA-3 (8B) Zero–shot attains a moderate *F1* (0.66) driven by recall-heavy predictions (0.77), yet at the cost of low precision (0.57) and modest accuracy (0.61). Zero–shot configurations like Phi-3 Medium (14B), Zero-shot, and LLaMA-2-7B Zero–shot also suffer from imbalanced outputs, often showing recall gains while underperforming on precision and accuracy. These inconsistencies, along with the need for frequent retraining in static models, limit their applicability for real-time, freshness-critical scenarios.

In contrast, our Dynamic-Switch model leverages two frozen generation models, one static and one recently updated, and a lightweight switch module that selects the temporally appropriate response. This setup eliminates the need for continuous model updates while ensuring timely and accurate predictions in evolving knowledge environments.

5.1. Qualitative Study

To further analyze the effectiveness of our proposed dynamic switching model framework, examples of our TaDD datatype (see Table 2). The dynamic model demonstrated its ability to identify subtle but substantial changes to facts, e.g., changes in numerical values, dates, and detailed factual descriptions. For instance, in the "Ben Hollioake" example, the model correctly identified temporal updates involving a substantial revision of historical facts. Such examples highlight the model's ability to accurately capture temporal change, which traditional static models might miss. Furthermore, qualitative evaluation showed the model's robust performance across domains such as sports, politics, technology, and general news. This domain adaptability reflects the model's effectiveness in dynamically and contextually selecting relevant models based on temporal alignment, as validated by human annotations and GPT-4.1 Mini evaluations. The qualitative analysis thus makes apparent the obvious advantage of taking temporal validity and semantic correctness into model selection criteria, enabling our approach to maintain higher factual accuracy and timeliness compared to static baselines.

5.2. Error Analysis

While our dynamic switching model demonstrates strong results, some cases reveal important shortcomings. First, we found that the model occasionally struggles with unit and quantity conversions. One example involved replacing "668m in offsets" with "668 hectares of land." The change in units was substantial despite the numerical value remaining the same. altered the meaning. The model incorrectly labelled this as not requiring an update. In the second case, there was an error involving named entities in the context of geographical references. A notable case involved replacing Greater Manchester with Greater London. Although this change is nuanced on the surface, it is not disregarded as a factual change. However, the model treated it as insignificant because of the similarity on the surface. Finally, we observed oversensitivity in some examples, where the model made purely stylistic edits. These false positives indicate that the model sometimes overreacts to

ID	Original Content (`old_content`)	Modified Content (`new_content`)	Original Label → Predicted
1	The directors denied any suggestion of wrongdoing or conflict of interest and said they made the appropriate declaration snow guardian australia can reveal that one of those directors – house holds interests in two other properties that sold a further 668m in offsets for developments in western sydney from 2017 to 2019.	The directors denied any suggestion of wrongdoing or conflict of interest and said they made the appropriate declaration snow guardian australia can reveal that one of those directors – house holds interests in two other properties that sold a further 668 hectares of land in the area for the development in western sydney from 2017 to 2019.	0 → 1
2	jersey's point pleasant beach on sunday with a united mission to pause offshore wind projects in response to recent whale deaths along the new york new jersey coast the gathering unfolded even as officials dispute the notion that the projects may be to blame for the dead whales a controversy.	jersey's point pleasant beach on sunday with a united mission to pause offshore wind projects in response to recent whale deaths along the new york new jersey coast the gathering unfolded even as officials from the state department of environmental protection, the department of conservation and the new jersey department of land use.	0 → 1
3	vodafone experienced a "catastrophic failure" kerslake found this caused significant stress and upset on the night to the families involved who were "reduced to a frantic search around the hospitals of greater manchester to find out more" vodafone.com	vodafone experienced a "catastrophic failure" kerslake found this caused significant stress and upset on the night to the families involved who were "reduced to a frantic search around the hospitals of greater london". To find out more vodafone.com	1 → 0

Table 7: Error analysis table highlighting content mismatches between old and new generations with incorrect model predictions. Red indicates the original mismatched phrase; blue indicates the predicted content.

rewording and minor additions that do not affect the factual content.

5.3. Human Evaluation

To assess the accuracy of model predictions regarding temporal freshness, we conducted a human evaluation of 1,000 randomly selected samples from the TaDD dataset. Each sample was labeled by our proposed model and seven baseline classifiers. Two expert annotators with backgrounds in NLP and linguistics independently reviewed the predicted freshness labels. The annotators were blind to model identities and instructed to rate each prediction on a scale of 1 to 5 under the dimension of Freshness[4].

Model	Freshness Score (1–5)
Ours (Dynamic-Switch)	**3.32**
GPT-4.1 Mini (Few-shot)	3.70
LLaMA-3 8B (Few-shot)	3.04
Phi-3 14B (Fine-tuned)	3.16
Mistral-7B (Fine-tuned)	3.02
LLaMA-3 8B (Fine-tuned)	3.18
GPT-2 (Fine-tuned)	2.48
BERT (Fine-tuned)	**2.04**

Table 8: Human evaluation of model-predicted freshness labels on 1000 TaDD samples.

Table 8 presents average freshness scores across models on 1000 TaDD samples. GPT-4.1 Mini (Few-shot) achieved the score (3.70), while the Dynamic-Switch model scored 3.32, outperforming several large fine-tuned models such as Mistral-7B (3.02) and LLaMA-3 8B (Few-shot) (3.04). Static baselines, such as GPT-2 (2.48) and BERT (2.04), performed poorly, highlighting their limitations in capturing temporal relevance. These results demonstrate the effectiveness and adaptability of our approach in mitigating temporal obsolescence. To ensure the reliability of these human evaluations, we computed inter-annotator agreement (IAA) between two annotators on the same 1,000-sample subset.

IAA Metric	Score
Cohen's Kappa (κ)	0.664
Raw Agreement (%)	83.20

Table 9: Inter-Annotator Agreement (IAA) between two human annotators on 1,000 TaDD samples. Cohen's Kappa indicates substantial agreement, confirming the reliability of the annotations.

As shown in Table 9, the resulting Cohen's Kappa (κ) of 0.664 and raw agreement of 83.2% indicate substantial agreement (Rau and Shih, 2021), confirming the consistency and reliability of the human annotation process.

6. Conclusion and Future Work

We introduced a dynamic model-switching framework that adaptively selects between a large, static model (Big_Model) and a small, frequently updated model (Small_Model) to enhance temporal accu-

[4]Freshness scores range from 1 (factually outdated or irrelevant) to 5 (fully temporally accurate and contextually fresh).

racy in LLMs. We have also curated a Temporally-aware Dynamic Dataset (TaDD) using our proposed framework. Our approach eliminates the need for continuous retraining while ensuring the generation of factually accurate and temporally coherent text. Experiments demonstrate significant improvements in Temporal Validity Assessment, annotated by humans with assistance from GPT-4.1 Mini. While effective, our framework primarily relies on Wikipedia and news datasets, which limit its domain diversity. Future work will focus on expanding data sources, evaluating real-time and domain-specific QA tasks, and exploring approaches such as retrieval-augmented generation to further enhance adaptability in evolving contexts.

7. Limitations

Although our proposed framework has shown interesting outcomes, it has some limitations. Our adaptive switching model dynamically chooses between temporal specialized models in order to keep the content relevant in time. Although this approach has the benefit of dynamically adapting in real time, it requires the existence of a historically trained model, in addition to a newly updated model, and this may not be feasible in terms of domain or deployment context. The framework implement a mechanism of model selection on the basis of both contextual and temporal information, however its performance may be affected by the degree of clarity and consistency of these two information types of input data. Additionally, although it avoids the need for continual retraining, it relies on the temporal relevance of the underlying models that can deteriorate over time in a highly dynamic information environment. This is a major drawback of the scheme that performance is still sensitive to the quality of model specialization and robustness switching mechanism, although its flexible, responsive architecture, which is designed to manage knowledge obsolescence with reduced maintenance overhead.

8. Acknowledgements

This research was partially supported by the Horizon Europe project *GenDAI* (Grant Agreement ID: 101182801) and by the ADAPT Research Centre at Munster Technological University. ADAPT is funded by Taighde Éireann – Research Ireland through the Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) via Grant 13/RC/2106_P2.

9. Bibliographical References

Axel Almquist and Adam Jatowt. 2019. Towards content expiry date determination: Predicting validity periods of sentences. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I*, volume 11437 of *Lecture Notes in Computer Science*, pages 86–101, Berlin, Heidelberg. Springer-Verlag.

Hsuvas Borkakoty and Luis Espinosa-Anke. 2023. Wikitide: A wikipedia-based timestamped definition pairs dataset.

Shuyang Cao and Lu Wang. 2022. Time-aware prompting for text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Taishi Hosokawa, Adam Jatowt, and Kazunari Sugiyama. 2024. Temporal validity reassessment: Commonsense reasoning about information obsoleteness. *Discovery Computing*, 27:4.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6750–6774.

Adam Jatowt, Ching-Man Au Yeung, and Katsumi Tanaka. 2013. Estimating document focus time. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 2273–2278, New York, NY, USA. Association for Computing Machinery.

Sunbowen Lee, Junting Zhou, Chang Ao, Kaige Li, Xinrun Du, Sirui He, Haihong Wu, Tianci Liu, Jiaheng Liu, Hamid Alinejad-Rokny, Min Yang, Yitao Liang, Zhoufutu Wen, and Shiwen Ni. 2025. Quantification of large language model distillation.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report. ArXiv preprint arXiv:2303.08774.

Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, Avishek Anand, and Adam Jatowt. 2025. It's high time: A survey of temporal information retrieval and question answering.

Dheeraj Rajagopal, Xuchao Zhang, Michael Gamon, Sujay Kumar Jauhar, Diyi Yang, and Eduard Hovy. 2022. One document, many revisions: A dataset for classification and description of edit intents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5517–5524, Marseille, France. European Language Resources Association.

Rishav Ranaut, Sriparna Saha, Adam Jatowt, and Manish Gupta. 2025. Text obsoleteness detection using large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2827–2831, Padua, Italy. Association for Computing Machinery.

Gerald Rau and Yu-Shan Shih. 2021. Evaluation of cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of English for Academic Purposes*, 53:101026.

Daivik Sojitra, Raghav Jain, Sriparna Saha, Adam Jatowt, and Manish Gupta. 2024. Timeline summarization in the era of llms. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2657–2661, Washington DC, USA. Association for Computing Machinery.

Alexander Spangher, Kung-Hsiang Huang, Hyundong Cho, and Jonathan May. 2024. Newsedits 2.0: Learning the intentions behind updating news.

Jonas Wallat, Abdelrahman Abdallah, Adam Jatowt, and Avishek Anand. 2025. A study into investigating temporal robustness of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15685–15705, Vienna, Austria. Association for Computational Linguistics.

Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. Temporal blind spots in large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, pages 683–692, Merida, Mexico. Association for Computing Machinery.

Jiapu Wang, Kai Sun, Linhao Luo, Wei Wei, Yongli Hu, Alan Wee-Chung Liew, Shirui Pan, and Baocai Yin. 2025. Large language models-guided dynamic adaptation for temporal knowledge graph reasoning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, pages Paper No. 269, 27 pages, Vancouver, BC, Canada. Curran Associates Inc.

Georg Wenzel and Adam Jatowt. 2023. An overview of temporal commonsense reasoning and acquisition.

Georg Wenzel and Adam Jatowt. 2024. Temporal validity change prediction. In *Findings of the Association for Computational Linguistics: ACL 2024*, ACL Findings, pages 1424–1446, Bangkok, Thailand. Association for Computational Linguistics.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Yipeng Zhang, Haitao Mi, and Helen Meng. 2025. Self-tuning: Instructing llms to effectively acquire new knowledge through self-teaching.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

A. Appendix

A.1. Comparative Performance Analysis

Table 10 has a detailed comparison of the zero-shots, few-shots, fine-tuned and LoRA-adapted models on the Combined Temporal Validity data. Relative gains or losses across metrics are emphasized by computing performance deltas (△) with respect to our DynamicSwitch model. Though several large finetuned models (e.g., Mistral-7B and Phi-3 Medium LoRA) are higher in absolute accuracy, these approaches require full or partial parameter updates. In context, our Dynamic-Switch framework maintains a competitive and balanced precision-recall framework that does not require retraining the underlying generators, and only trains a lightweight 120M-parameter switch module. The table also compares the parameter counts for training and inference, highlighting the efficiency trade-off between static fine-tuning and our dynamic switching method.

A.2. Annotation Guidelines

To ensure reliable temporal validity labeling, we followed a structured two-stage annotation procedure as the one outlined in Section 4. Preliminary draft labels were created as an assistive tool with GPT-4.1 Mini and final decisions were taken solely by human annotators. The annotators were asked to mark the sentence pairs into two categories:

- **Temporal Update (Label = 1):** In the case of substantial changes to facts, such as, changes in numerical values, dates, scores, quantities, named entities, status updates, or time-sensitive information that modify the factual validity of the statement.
- **Semantic Update (Label = 0):** Minor editorial changes like rewording, stylistic modifications, punctuation, paraphrasing, or elaborations, which do not change the factual meaning.

It was made clear to the annotators that attention was to be paid to the factual shifts and not lexical similarity. In case of ambiguity, the focus was on the presence of new time sensitive information in the update which would be used to influence temporal accuracy. Cohen Kappa was used to assess the level of inter-annotator agreement and the results were 0.664 which was high, thus, the reliability of the annotations is confirmed.

Model	Accuracy (△)	Precision (△)	Recall (△)	F1-Score (△)	Active Params (Train / Infer)
GPT-4.1 Mini (Few-shot)	0.78 (↑20%)	0.65 (↓2%)	0.72 (↑11%)	0.68 (↑4%)	0 / 14B
GPT-4.1 Mini (Zero-shot)	0.75 (↑15%)	0.60 (↓9%)	0.70 (↑8%)	0.65 (↓1%)	0 / 14B
LLaMA-3 (8B) Few-shot	0.65 (=0%)	0.70 (↑6%)	0.93 (↑43%)	0.80 (↑22%)	0 / 8B
LLaMA-3 (8B) Zero-shot	0.61 (↓6%)	0.57 (↓14%)	0.77 (↑18%)	0.66 (↑1%)	0 / 8B
LLaMA-3 (8B) LoRA-Fine-Tuned	**0.69** (↑4%)	**0.67** (↑1%)	**0.79** (↑14%)	**0.72** (↑7%)	**140M / 8B**
LLaMA-2-7B Zero-shot	0.58 (↓11%)	0.57 (↓14%)	0.59 (↓9%)	0.58 (↓11%)	0 / 7B
Mistral-7B Zero-shot	0.48 (↓26%)	0.47 (↓29%)	0.60 (↓8%)	0.53 (↓19%)	0 / 7B
Mistral-7B Fine-Tuned	0.75 (↑15%)	0.72 (↑9%)	0.78 (↑20%)	0.75 (↑15%)	7B / 7B
Mistral-7B LoRA-Fine-Tuned	**0.68** (↑3%)	**0.64** (↑2%)	**0.70** (↑5%)	**0.67** (=0%)	**150M / 7B**
GPT-2 Fine-Tuned	0.72 (↑11%)	0.75 (↑14%)	0.70 (↑8%)	0.72 (↑10%)	124M / 124M
BART Fine-Tuned	0.69 (↑6%)	0.62 (↓6%)	0.64 (↓2%)	0.63 (↓4%)	139M / 139M
Our Model (Dynamic-Switch)	**0.65** (=0%)	**0.66** (=0%)	**0.65** (=0%)	**0.65** (=0%)	**120M / (14B + 3.8B)**
Phi-3 Mini Fine-Tuned	0.70 (↑8%)	0.62 (↓6%)	0.74 (↑14%)	0.67 (↑2%)	3.8B / 3.8B
Phi-3 Mini Zero-shot	0.55 (↓15%)	0.50 (↓24%)	0.79 (↑22%)	0.61 (↓6%)	0 / 3.8B
Phi-3 Medium (14B) Few-shot	0.67 (↑3%)	0.65 (↓2%)	0.73 (↑12%)	0.69 (↑5%)	0 / 14B
Phi-3 Medium (14B) Zero-shot	0.65 (=0%)	0.60 (↓9%)	0.72 (↑11%)	0.65 (=0%)	0 / 14B
Phi-3 Medium (14B) LoRA-Fine-Tuned	**0.71** (↑6%)	**0.66** (↑4%)	**0.75** (↑1%)	**0.70** (↑3%)	**145M / 14B**

Table 10: Performance comparison on the Combined Temporal Validity dataset with LoRA variants added. Colored arrows (↑, ↓, =) indicate percentage change relative to our Dynamic-Switch model. **Active Params (Train / Infer)** shows parameters used for training and inference, respectively; LoRA rows report only trainable adapter params / full inference model size.