Continual Learning and Memory Models Aksel Kretsinger-Walters, adk2164
Temporal Generalization in Language Models adk2164@columbia.edu
COMS 6998: Spring 2026 **Due:** 02/04/2026 11:59 PM ET

Mind the Gap: Temporal Generalization in Language Models

Motivation. Language is non-stationary and open-ended, yet modern language models are trained and evaluated on static datasets with overlapping time periods, obscuring their ability to generalize to future data and motivating a systematic study of temporal generalization.

Static Benchmarking Limitations. Standard evaluation practices focus on perplexity and static downstream benchmarks, which fail to measure performance on future utterances and increase the risk of test-set contamination through temporal overlap.

Research Questions. The lecture investigates how much static evaluation overestimates model performance, how temporal degradation manifests across tasks and linguistic categories, and whether scaling or adaptive methods can mitigate performance decay.

Experimental Setup. Models are evaluated on time-stratified test sets drawn from future data relative to training, with controlled baselines that inject future documents into training to isolate the effect of temporal mismatch.

Datasets and Model. Experiments use temporally annotated corpora including WMT, CustomNews, and arXiv, with evaluation spanning monthly test sets from 2018–2019 and a Transformer-XL model trained with a fixed context length.

Temporal Degradation. Results show that perplexity increases monotonically as test data moves further beyond the training period, demonstrating consistent temporal degradation across domains.

Where Degradation Occurs. Errors grow disproportionately for proper nouns, numbers, and time-sensitive topics such as politics and economics, while syntactic categories and reading comprehension tasks are less affected.

Effect of Scale. Larger models reduce but do not eliminate temporal degradation, indicating that scale alone cannot solve the problem of outdated knowledge.

Dynamic Evaluation. Online gradient updates at inference time improve perplexity on future data but introduce catastrophic forgetting, highlighting the need for memory-aware or continual learning mechanisms.

Key Takeaways. Temporal generalization should be treated as a first-class evaluation axis, and future language models must incorporate adaptive, retrieval-based, or memory-driven mechanisms to remain accurate in a dynamic world.