universität innsbruck

When Do Answers Change? Estimating Question Recency Demands in QA with Multi-Events

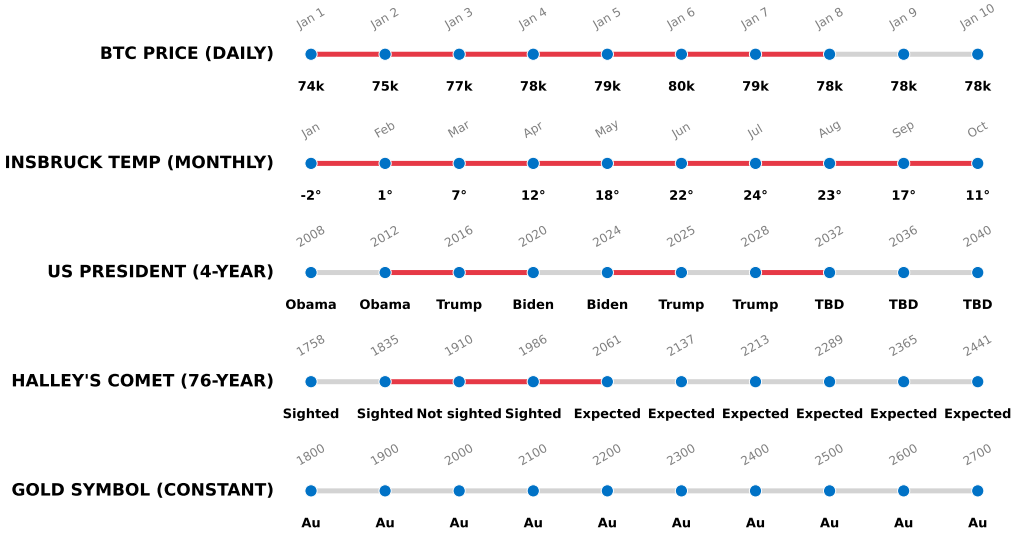
Recency-aware QA dataset, pipeline, and results

Peter Schulze Fabian Stiewe

Agenda

- **Motivation and Prior Work**
- Dataset Creation
- Testing Pipeline
- Results
- Fine-Tuning
- Conclusion and Outlook

Motivation


BTC PRICE (DAILY)
Jan 1 Jan 2 Jan 3 Jan 4 Jan 5 Jan 6 Jan 7 Jan 8 Jan 9 Jan 10
74k 75k 77k 78k 79k 80k 79k 78k 78k 78k
INSBRUCK TEMP (MONTHLY)
Jan Feb Mar Apr May Jun Jul Aug Sep Oct
-2° 1° 7° 12° 18° 22° 24° 23° 17° 11°
US PRESIDENT (4-YEAR)
2008 2012 2016 2020 2024 2025 2028 2032 2036 2040
Obama Obama Trump Biden Biden Trump Trump TBD TBD TBD
HALLEY'S COMET (76-YEAR)
1758 1835 1910 1986 2061 2137 2213 2289 2365 2441
Sighted Sighted Not sighted Sighted Expected Expected Expected Expected Expected Expected
GOLD SYMBOL (CONSTANT)
1800 1900 2000 2100 2200 2300 2400 2500 2600 2700
Au Au Au Au Au Au Au Au Au Au

Why recency demand?

- Answer to questions change
- **Recency demand**: refresh rate; to keep answer up-to-date

Goal

Build datasets that stress temporal reasoning and evaluate **LLM** performance on recency demand.

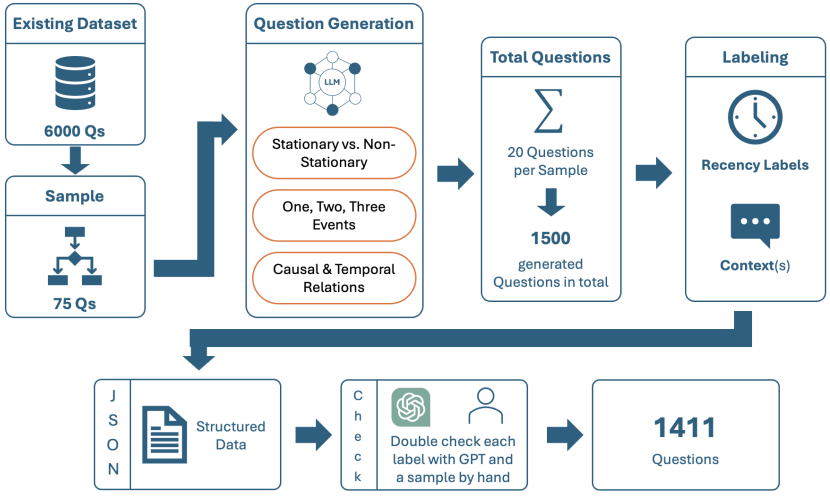
Related datasets

Dataset	Creation	Knowledge	KC	Recency	Multi-Event	#Q
TimeQA	Templ.	Wikipedia	✗	✗	✗	20 000
StreamingQA	Man.+Gen.	News	✓	✗	✗	410 000
RealTimeQA	News	News	✓	✗	✗	5000
PATQA	Templ.	Wikipedia	✓	✗	✗	6172
FreshQA	Manual	Web	✓	✗	✗	600
RecencyQA (orig.)	Man.+Gen.	Wiki	✓	✓	✗	6115
RecencyQA-Multi (ours)	Man.+Gen.	RecencyQA+LLM	✗	✓	✓	1411

Agenda

- Motivation and Prior Work
- ▶ **Dataset Creation**
- Testing Pipeline
- Results
- Fine-Tuning
- Conclusion and Outlook

Pipeline overview



Question taxonomy

- Recency Classes (12)
- Stationarity
- Event structure
- Inter-event relation
- Contextual conditions

12 Recency Classes

Class	Time to change
An-Hour	Within an hour
A-Few-Hours	Within a few hours
A-Day	Within a day
A-Few-Days	Within a few days
A-Week	Within a week
A-Few-Weeks	Within a few weeks
A-Month	Within a month
A-Few-Months	Within a few months
A-Year	Within a year
A-Few-Years	Within a few years
Many-Years	After many years
Never	No change

Prompt Families

St.	#Ev.	Rel.	Prompt template
S	1	–	Stationary single
S	2	C	Stationary 2-event causal
S	2	T	Stationary 2-event temporal
S	3	C	Stationary 3-event causal
S	3	T	Stationary 3-event temporal
NS	1	–	Non-stationary single
NS	2	C	Non-stationary 2-event causal
NS	2	T	Non-stationary 2-event temporal
NS	3	C	Non-stationary 3-event causal
NS	3	T	Non-stationary 3-event temporal

12 Recency classes

Class	Time to change	Example
An-Hour	Within an hour	Current Apple stock price
A-Few-Hours	Within a few hours	Traffic on A9 highway
A-Day	Within a day	Today’s weather in Munich
A-Few-Days	Within a few days	Trending movies on Netflix
A-Week	Within a week	Top Billboard songs this week
A-Few-Weeks	Within a few weeks	FIFA ranking of Germany
A-Month	Within a month	Unemployment rate in Italy
A-Few-Months	Within a few months	Inflation rate in Eurozone
A-Year	Within a year	Current Java version
A-Few-Years	Within a few years	President of the US
Many-Years	After many years	Population of Germany
Never	No change	Chemical symbol for gold

Question taxonomy

- Recency Classes (12)
- Stationarity
- Event structure
- Inter-event relation
- Contextual conditions

12 Recency Classes

Class	Time to change
An-Hour	Within an hour
A-Few-Hours	Within a few hours
A-Day	Within a day
A-Few-Days	Within a few days
A-Week	Within a week
A-Few-Weeks	Within a few weeks
A-Month	Within a month
A-Few-Months	Within a few months
A-Year	Within a year
A-Few-Years	Within a few years
Many-Years	After many years
Never	No change

Prompt Families

St.	#Ev.	Rel.	Prompt template
S	1	–	Stationary single
S	2	C	Stationary 2-event causal
S	2	T	Stationary 2-event temporal
S	3	C	Stationary 3-event causal
S	3	T	Stationary 3-event temporal
NS	1	–	Non-stationary single
NS	2	C	Non-stationary 2-event causal
NS	2	T	Non-stationary 2-event temporal
NS	3	C	Non-stationary 3-event causal
NS	3	T	Non-stationary 3-event temporal

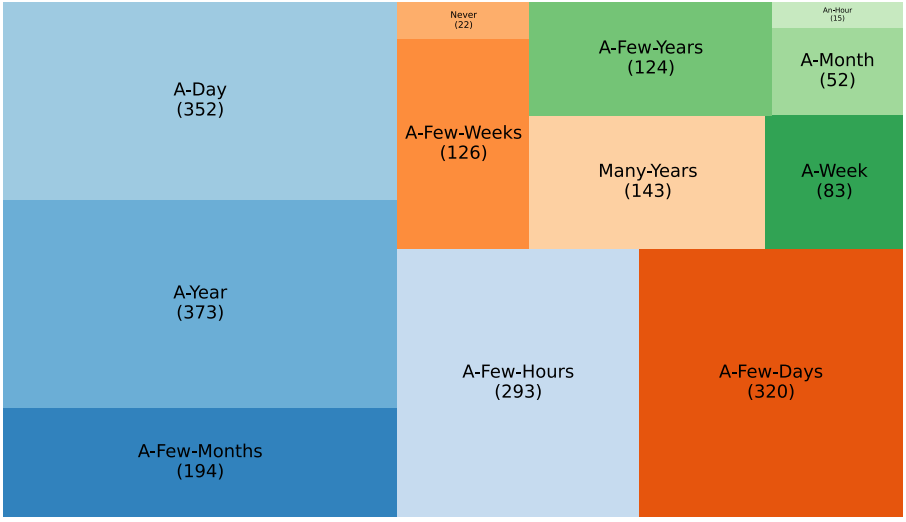
Prompt families for generation

St.	#Ev.	Rel.	Prompt template
S	1	–	Stationary single
S	2	C	Stationary 2-event causal
S	2	T	Stationary 2-event temporal
S	3	C	Stationary 3-event causal
S	3	T	Stationary 3-event temporal
NS	1	–	Non-stationary single
NS	2	C	Non-stationary 2-event causal
NS	2	T	Non-stationary 2-event temporal
NS	3	C	Non-stationary 3-event causal
NS	3	T	Non-stationary 3-event temporal

Dataset statistics

Metric	Value
Total questions	1411
Stationary / Non-stationary	725 / 686
Single / Two / Three events	286 / 555 / 570
Multi-event (causal / temporal)	565 / 560
Avg. question length (tokens)	22.2
Avg. context length 1 / 2	11.2 / 12.6
Total recency labels	2097
Top labels	A-Year, A-Day, A-Few-Days

Recency label distribution



Agenda

- Motivation and Prior Work
- Dataset Creation
- ▶ **Testing Pipeline**
- Results
- Fine-Tuning
- Conclusion and Outlook

Testing pipeline

- Flatten dataset: each question-context-label pair becomes one instance.
- Testing on three models:
  - *Kimi-K2-Instruct-0905*
  - *Qwen2.5-72B-Instruct-Turbo*
  - *DeepSeek-V3*
- Collect predictions per model into `JSONL`
- Summarize accuracy and tolerant (± 1 label) accuracy.
- Slice metrics:
  - Stationary vs. non-stationary
  - Single-event vs. multi-event
  - Causal vs. temporal-only

Agenda

- Motivation and Prior Work
- Dataset Creation
- Testing Pipeline
- ▶ **Results**
- Fine-Tuning
- Conclusion and Outlook

Model overview

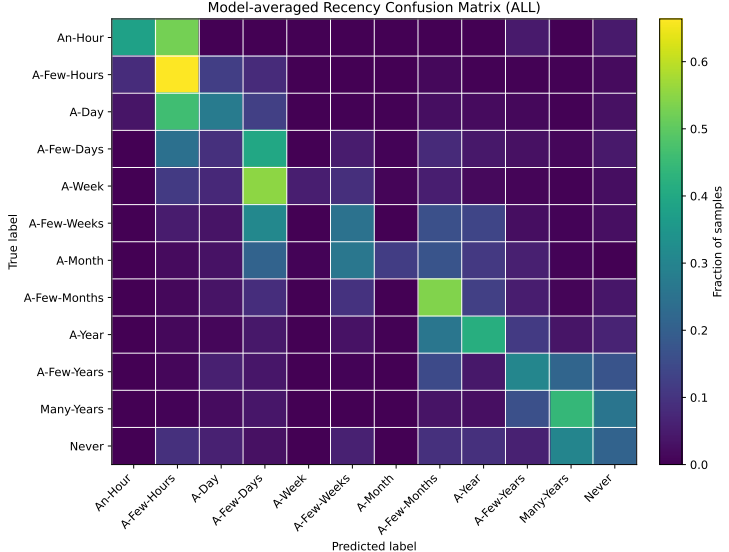
Model	Overall		St.		Non-St.		# Events Acc			Multi-event Acc	
	Acc	Tol.	Acc	Tol.	Acc	Tol.	1	2	3	C	T
Kimi-K2-0905	33.0	62.2	42.1	66.3	28.3	60.0	35.0	31.1	33.6	34.9	29.8
Qwen2.5-72B	45.6	77.4	62.8	81.8	36.7	75.1	45.4	44.6	46.6	47.8	43.5
DeepSeek-V3	40.8	71.9	52.2	73.2	34.9	71.1	44.9	38.1	41.1	43.9	35.6
Average	39.8	70.5	52.4	73.8	33.3	68.7	41.7	37.9	40.5	42.2	36.3

Findings by slice

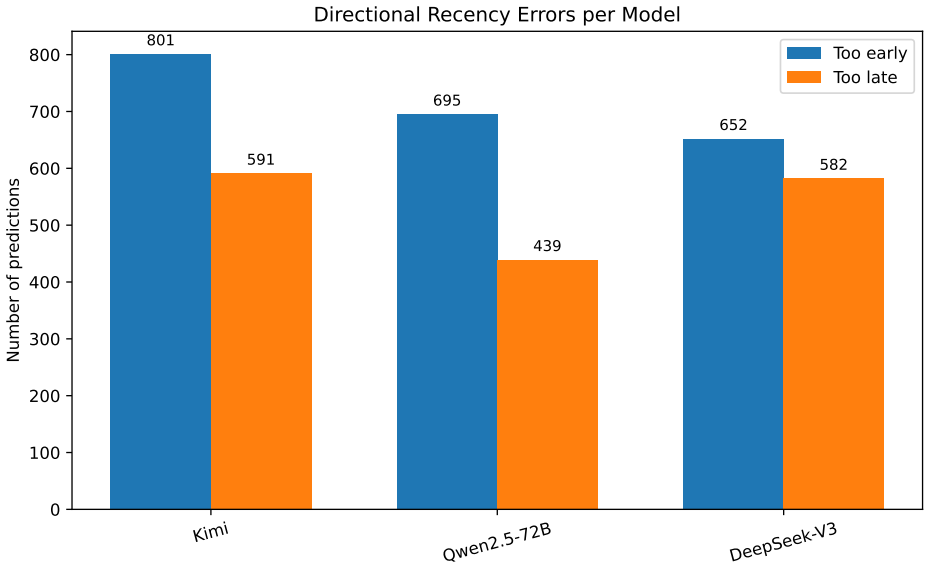
- Biggest drop: stationary $\rightarrow$ non-stationary (-13 to -26 pts)
- Multi-event hurts slightly ($\sim$)4 pts)
- Causal is easier than temporal-only
- Tolerant scores $+20$ to $+30$ pts $\rightarrow$ many near-misses

Model	Overall		St.		Non-St.		# Events Acc			Multi-event Acc	
	Acc	Tol.	Acc	Tol.	Acc	Tol.	1	2	3	C	T
Kimi-K2-0905	33.0	62.2	42.1	66.3	28.3	60.0	35.0	31.1	33.6	34.9	29.8
Qwen2.5-72B	45.6	77.4	62.8	81.8	36.7	75.1	45.4	44.6	46.6	47.8	43.5
DeepSeek-V3	40.8	71.9	52.2	73.2	34.9	71.1	44.9	38.1	41.1	43.9	35.6
Average	39.8	70.5	52.4	73.8	33.3	68.7	41.7	37.9	40.5	42.2	36.3

Confusion across recency labels



Directional errors

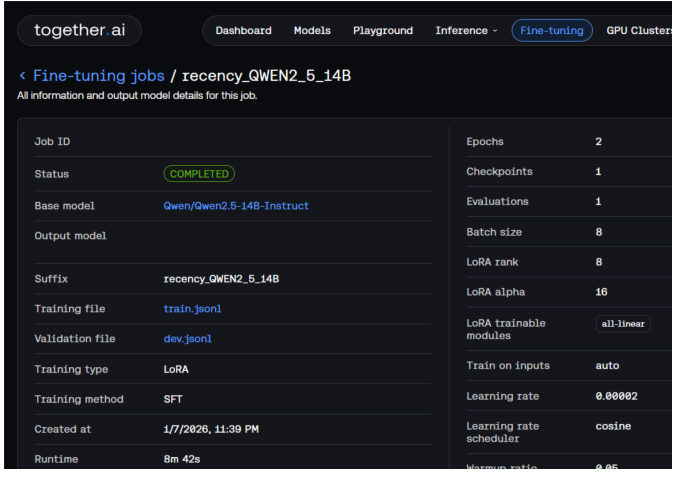


Agenda

- Motivation and Prior Work
- Dataset Creation
- Testing Pipeline
- Results
- ▶ **Fine-Tuning**
- Conclusion and Outlook

Fine-tuning setup

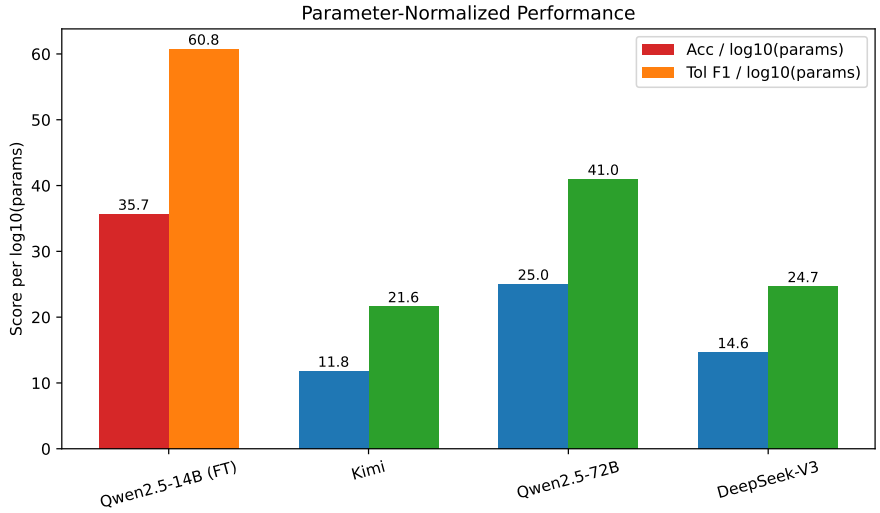
- Smaller Qwen model (14B)
- Split Dataset 70/15/15
- Finetuned via Together AI
- Compared to previous models



Fine-tuning results (reduced test set)

Model	Overall		St.		Non-St.		Single-event		Multi-event	
	Acc	Tol.	Acc	Tol.	Acc	Tol.	Acc	Tol.	Acc	Tol.
Qwen2.5-14B (FT)	40.9	69.6	55.1	72.9	33.5	68.0	41.5	78.5	40.7	67.3
Kimi	35.5	64.9	46.3	70.4	29.8	62.0	29.7	65.6	36.9	64.7
Qwen2.5-72B	46.5	76.1	63.9	82.4	37.4	72.8	46.2	87.7	46.6	73.1
DeepSeek-V3	41.4	69.7	50.9	71.3	36.4	68.9	38.5	70.8	42.2	69.5

Parameter efficiency



Agenda

- Motivation and Prior Work
- Dataset Creation
- Testing Pipeline
- Results
- Fine-Tuning
- ▶ **Conclusion and Outlook**

Key takeaways

- 1411 questions, 12 classes, controlled stationarity & events
- Tolerant >70%, strict accuracy remains low
- Hard cases: non-stationary, mid-horizon; “too-early” bias

Next steps

- Soft labels (recency distributions) to capture uncertainty
- Parallel LLM labeling and majority voting for better quality
- Human labeling vs. LLM-based

Questions?

Human labeling

https://recency-labeling-page.vercel.app/

Labeling and verification

- Label recency classes:
  - Stationary: single label + context
  - Non-Stationary: two labels + contexts
- Use *Llama-3.3-70B-Instruct-Turbo* for labeling
- Enforce structured JSON
- LLM screening (*GPT-5.1 Codex Max*) for label-question consistency
- Manual audit of flagged items; remove or fix ambiguous cases
  - Sample 75 items for manual verification