

Can Structural Cues Save LLMs?

Evaluating Language Models in Massive Document Streams

Yukyung Lee,¹ Yebin Lim,^{2,*} Woojun Jung,^{2,*} Wonjun Choi,² and Susik Yoon,²

¹Boston University, ²Korea University

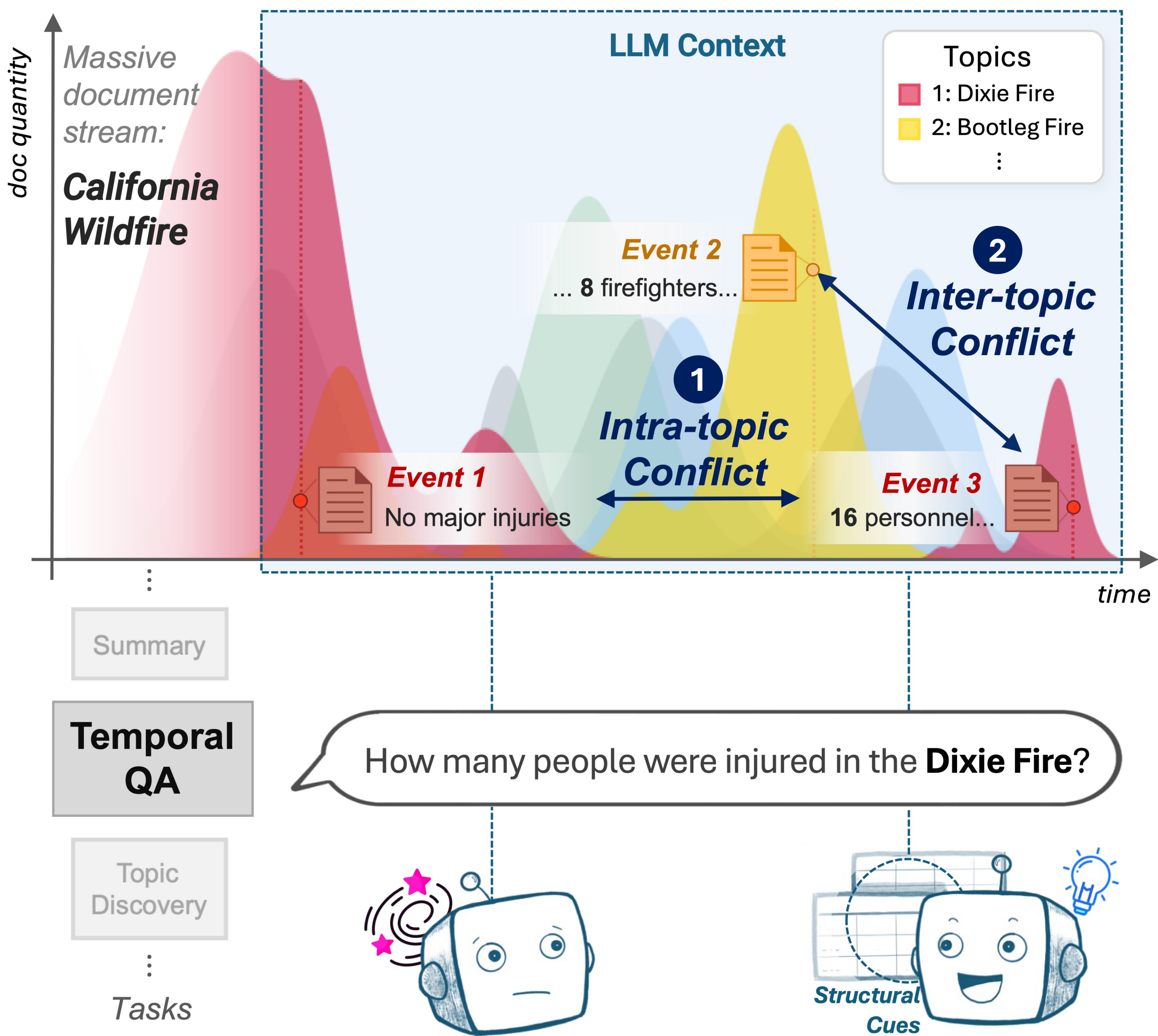


1. Background

Challenges in Massive Document Streams

Streaming Setting

Documents from multiple concurrent events arrive mixed together over time within a fixed context window.



Two Key Conflicts

- 1) **Intra-topic Conflict:** Accumulated older information biases models away from recent updates.

2) **Inter-topic Conflict:** Models struggle to distinguish facts across related topics.

Core Challenge

These conflicts require LLMs to *organize scattered information* by topic and time *while reasoning over it*.

Organization Bottleneck

Reasoning Bottleneck

Existing Benchmark Gap

Existing benchmarks focus on static snapshots or single events, and do not evaluate models under these conflicts.

2. Motivation

Do failures stem from difficulties in organizing scattered information, or from reasoning over it?

Structural Cues as a Diagnostic Probe

- Raw Input:

Documents from multiple topics are mixed together.
- Cued Input:

The same documents with key facts organized by event.
- structural cue = (People, Location, Result, Event Attributes)

Diagnostic Measures



3. StreamBench

6 Stories 200 Topics 605 Events 15,354 Docs					
Story	Contents	Duration	Topic	Event	Docs
A	California Fire	Jan'24–Nov'25	32	113	768
B	South Korea Martial Law	Jun'24–Nov'25	15	108	1,135
C	60th US Presidential Election	Jan'24–Nov'25	32	111	724
D	Summer Olympics	Apr'16–Dec'16	20	35	977
E	Israel-Palestine Conflict	Jan'16–Dec'16	13	45	387
F	58th US Presidential Election	Jan'16–Dec'16	88	193	11,363

Benchmark Setting

- Documents arrive in time order, without topic or event labels.

$k \in \{1, 3, 5, 10\}$ docs per event controls stream volume; larger k increases intra- and inter-topic conflicts

Tasks

- Topic Clustering:

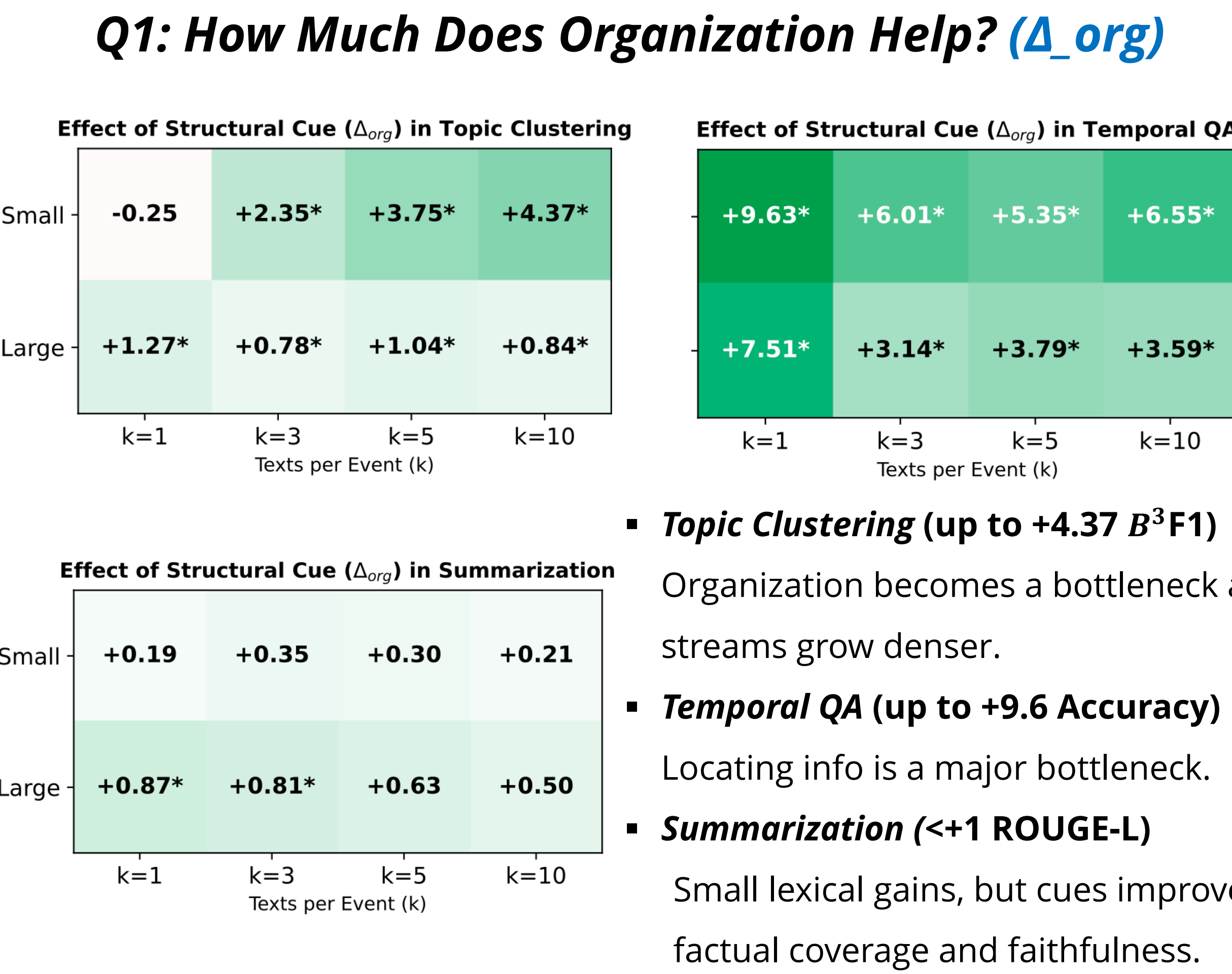
Assign each document to an existing or new topic
- Temporal QA:

Answer questions using past documents and prioritize recent information.
- Summarization:

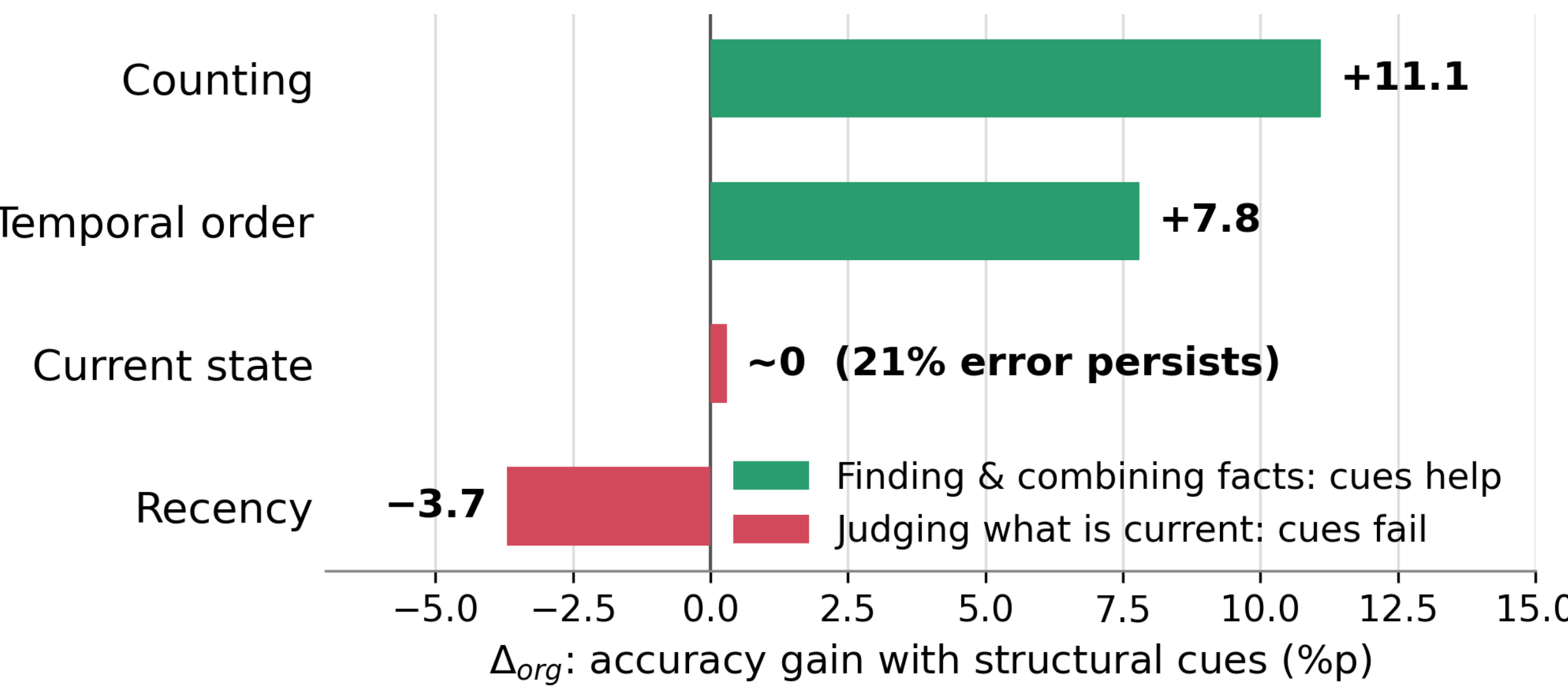
One summary covering all topics in the window

4. Findings

- We evaluate 7 instruction-tuned LLMs, ranging from 1B to 123B.



Task	Q2: What Remains Difficult ? (Δ_{gap})	
	What Organization Helps	What Remains Difficult
Topic Clustering	Separating mixed events	Precise boundary detection
Temporal QA	Locating and aggregating evidence	Current-state tracking and recency judgment
Summarization	Factual coverage and faithfulness	Coherent information integration



Structural cues help LLMs organize and locate information, but temporal reasoning and coherent integration remain difficult.