RETRIEVAL APOPHENIA: HOW GROWING CANDIDATE POOLS MAKE NEW EVENTS LOOK FAMILIAR

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming event induction asks whether an arriving document belongs to an event already in memory or starts a new one. Most systems answer this question by retrieving historical candidates, taking the candidate with the highest compatibility score, and comparing that score with a fixed threshold. We show that this maximum rule fails in a specific and predictable way: as memory grows, the maximum is taken over more candidates, making a genuinely new event increasingly appear familiar. We call this phenomenon candidate-pool false familiarity. Prior work has studied threshold drift in streams and score inflation in large candidate galleries, but these effects have not been connected within streaming event memory, where an erroneous merge changes the state against which subsequent documents are judged. We measure this effect on annotated non-coreferent candidates and distinguish between two possible explanations. If the drift were merely an uncalibrated score scale, normalising each query by its candidate pool should eliminate it. However, this removes only a small portion of the effect, indicating that the decision must examine the candidates themselves rather than rely on a single score. We therefore propose C2E-JUDGE, which keeps the retriever, candidate budget, and memory fixed while allowing a language-model judge to inspect the entire bounded pool. It separates the two error types: adding candidates improves repeated-event coverage with nearly no increase in false merges, whereas replacing score thresholding with whole-pool judging removes most false merges. This visibility effect persists under a stricter prompt, a second judge endpoint, and a second dataset.

1 INTRODUCTION

Streaming event induction is harder than it looks, because memory itself creates the problem. As a system accumulates history, it acquires both the evidence needed to recognise repeated events and many more chances for a new document to look familiar. The decision is irreversible: once a document is merged into the wrong event, that error propagates into every later retrieval. This setting underlies cross-document event coreference and online new-event detection, where a system must decide, for each arriving document, whether it joins an existing event or starts a new one (Allan et al., 1998; Eirew et al., 2021).

Most systems make this decision with the same three-stage pipeline. They retrieve historical candidates, score each candidate against the incoming document, and accept the best match if its score passes a threshold (Allaway et al., 2021; Logan IV et al., 2021). Recent work improves the pipeline by changing the representation or the inference path: graph-structured event representations, collaborative inference between mentions, temporal and discourse constraints, and richer benchmarks (Ahmed et al., 2024a; Min et al., 2024; Chen et al., 2025; 2026; Ahmed et al., 2024b; Zhao et al., 2025). Adaptive-threshold methods tackle the threshold directly by letting it evolve with the stream (Chen et al., 2009). All of them, however, still reduce the retrieved candidates to one number—the maximum compatibility score—and compare that number with a fixed cutoff.

That reduction is where the trouble lies, because a maximum depends on how many candidates it is taken over. The same pressure is well understood in large-gallery identification, open-set recognition, and selective prediction (Baveja et al., 2010; Bendale & Boult, 2016; Jin & Candès, 2023; Fisch et al., 2022): the more alternatives are available, the more likely it is that an implausible one

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

scores highly. Streaming event induction is a natural place for this effect to matter, because there the candidate pool is not a design choice—it grows with the stream. Yet we are not aware of prior work that isolates the effect inside a streaming event memory, where a false match changes the state that later decisions are made against.

We isolate the effect with a diagnostic that removes every confound we can control. Instead of reading a single clustering score, we compare each event's first mention only against earlier candidates that are annotated as different events, so no compared pair is coreferent and the comparison is auditable. The maximum score over these non-coreferent candidates rises steadily as the pool grows, and a fixed threshold therefore becomes less and less selective. We call this candidate-pool false familiarity.

We then ask what the rise actually means, and the answer changes what the remedy must be. Two explanations are possible: either the scores sit on an uncalibrated scale that grows with the pool, in which case rescaling the threshold would fix it, or the non-coreferent candidates are genuinely close in meaning, in which case no rescaling can help. We separate the two by replacing the raw maximum with a pool-relative extremeness score computed from each query's own candidate distribution. Normalising this way removes only part of the drift, so the phenomenon is not a calibration artifact: a system has to look at the candidates, not only at the score.

This finding points to the method we propose. C2E-JUDGE leaves the retriever, the candidate budget, and the streaming memory untouched, and instead lets a language-model judge read the whole bounded candidate set, returning either a supplied candidate identifier or UNSEEN. Language models have been used as pairwise judges for event identity (Zhao et al., 2023), but our setting differs in two ways: the judge must choose from a growing memory of events or reject every visible candidate, and the candidate set it reads is itself the object of study. We therefore separate visibility from model capacity with a controlled sweep that holds the judge, prompt, retriever, and memory fixed and varies only the number of visible candidates. Figure 1 summarizes the setting, the diagnosis, and our method.

Our contributions are:

1. We identify and quantify candidate-pool false familiarity in streaming event induction, using a non-coreferent diagnostic, a paired same-query analysis, and a pool-relative normalisation control that shows most of the drift is not a rescaling artifact.
2. We turn the diagnostic into a controlled dose–response experiment: with the same judge and retrieval budget, increasing the visible candidate count raises repeated-event coverage at near-zero false-merge risk, with per-step uncertainty intervals. Visibility and the decision model contribute to different error types, and we report them separately.
3. We provide the candidate-conditioned event-judge interface and evaluate it under an online protocol that never uses labels or future documents, testing the finding with a stricter prompt, a second judge endpoint, and a second dataset.

2 RELATED WORK

Our work sits between three literatures, and it borrows a warning from a fourth. Streaming event coreference and online detection define the task and the pipeline we study. Selective prediction and large-gallery identification explain why a maximum over a growing candidate set behaves the way it does. Open-set recognition and event-rejection work share our match-versus-new decision but not our setting. The paragraphs below take these in turn and state what each one does not cover.

Cross-document event coreference provides the task but assumes a fixed candidate set. WEC supplies a large cross-document event-coreference collection derived from Wikipedia, and we use its identity annotations for our primary diagnostic (Eirew et al., 2021). Sequential cross-document coreference studies the incremental setting in which documents arrive over time and earlier decisions constrain later ones (Allaway et al., 2021), and streaming entity benchmarks study the same bounded-memory regime (Logan IV et al., 2021). More recent systems improve the representation or the inference path: graph-structured events and multi-hop decisions (Ahmed et al., 2024a), collaborative inference between mentions (Min et al., 2024), richer benchmarks (Wang et al., 2022;

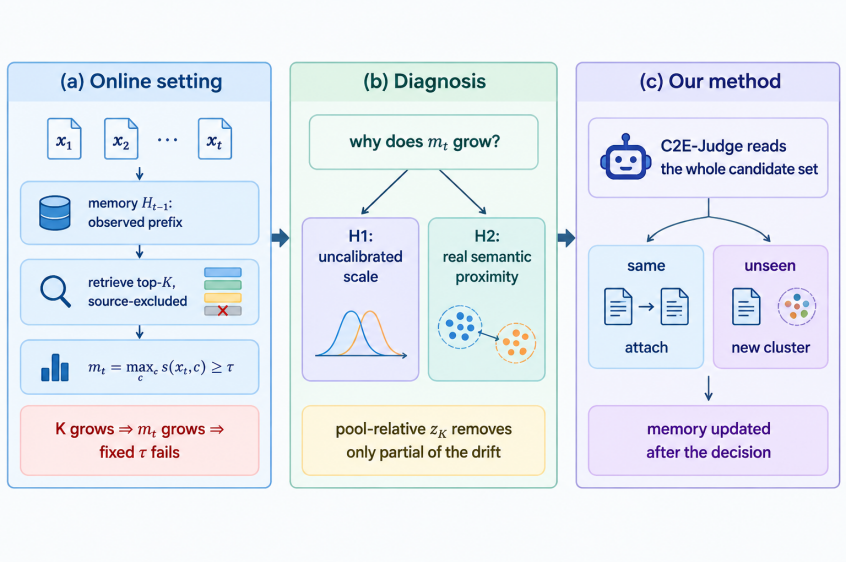


Figure 1: Overview of the problem, the diagnosis, and our method. (a) In the online setting, documents arrive one at a time and only the observed prefix is kept. The incoming document is scored against the retrieved candidates by the maximum score m_t, which grows with the pool size K, so a fixed threshold τ loses selectivity. (b) This growth has two explanations: an uncalibrated score scale, which a pool-relative score z_K would remove, and genuine semantic proximity, which it would not. Normalisation removes only about 24% of the drift, so a system must read the candidates themselves. (c) C2E-JUDGE instead lets one judge read the whole candidate set and return SAME or UNSEEN; memory is updated after the decision.

Ahmed et al., 2024b; Zhao et al., 2025), and discourse or temporal constraints added before resolution (Chen et al., 2025; 2026). xCoRe is a strong cross-context system that combines within-context resolution with supervised cluster merging (Martinelli et al., 2025); we cite it as an architectural reference rather than placing its offline CoNLL score on our Pareto curve, because it consumes cross-context inputs and emits document clusters, whereas our protocol makes one irreversible decision per arriving mention. This literature does not isolate how the decision statistic itself behaves as the candidate set grows: the systems are compared at a fixed operating point, and the pool size is treated as a retrieval setting rather than as a variable that changes the meaning of the score.

Language-model judges have been used for event identity, but for isolated pairs. GPT-4 has been evaluated as a zero-shot judge of whether two decontextualized cross-document event mentions corefer (Zhao et al., 2023). That work measures agreement with human annotation on isolated pairs. Our judge faces a different decision: it must either select one candidate from a growing memory or reject every candidate it sees, and the membership of the candidate set is the variable we study rather than a fixed input.

Selective prediction and large-gallery identification explain the mechanism but study different objects. Conformal selection calibrates prediction sets or p-values when only selected items are acted upon (Jin & Candès, 2023); conformal risk control extends this to bounded losses (Angelopoulos et al., 2024); and later work combines selection with risk control or controls selection errors across many tests (Xu et al., 2025; Wang et al., 2024; Fisch et al., 2022). We borrow their general warning that selecting on a score changes what the score means, but our object differs in two ways. First, one incoming mention competes against a memory of historical event clusters, and the decision uses the maximum over that memory. Second, we make no distribution-free or finite-sample guarantee; our claim is about observed behaviour under a streaming protocol. Biometric identification provides the closest precedent for the mechanism itself, showing that maximum impostor scores grow with gallery size (Baveja et al., 2010). We do not claim that pressure is new. What we add is its consequence inside a streaming event memory, where a false match alters the

state against which later documents are judged, together with the risk–coverage measurement that follows.

Online detection and adaptive thresholds address drift, but not pool size. The topic-detection literature established the online decision of whether a document starts a new story or continues an existing one (Allan et al., 1998), and adaptive-threshold methods adjust the cutoff as an event evolves, for instance by modelling its life profile (Chen et al., 2009). Together they show that one static threshold is rarely adequate in a stream. The source of instability we study is different: the number of competing historical candidates changes the distribution of the maximum even when the incoming documents and the pair scorer are fixed. The conditioning variable is therefore the candidate pool itself.

Open-set and rejection work shares the decision, not the setting. Open-set recognition, open-world learning, and hubness all deal with unknown classes, weak tail evidence, or over-frequent neighbours (Jafarzadeh et al., 2022; Bendale & Boult, 2016; Nielsen & Hansen, 2024). They share the match-versus-new-event tension with our task, but they do not study a candidate pool that grows over time and is itself the cause of the failure. Our targeted review found no work that combines a growing candidate pool, an explicit non-coreferent maximum-score diagnostic, and a candidate-conditioned remedy under one fixed retrieval budget. This is a scoped distinction about the task, not a claim that the ingredients are new.

3 Preliminaries and Problem Formulation

This section fixes the setting and the notation. We describe the online streaming protocol, define the decision statistic that we will criticise, and define the two error rates we report throughout. Throughout the paper "causal" is not used in the causal-inference sense: it only means that a decision may use the observed prefix alone and cannot be revised later.

3.1 Online streaming event induction

The setting is a single pass over a stream with no access to the future. At time t, an event-bearing document x_t and its supplied mention e_t arrive. The system holds a memory $\mathcal{H}_{t-1}$ of historical reports and the predicted event clusters they belong to. It may use only this prefix and the current document; the current report enters memory after the decision, so no future document and no label is available when the choice is made. This one-pass constraint is what makes an early mistake persistent, since the wrong cluster is then available as evidence for every later decision.

3.2 The standard decision statistic is pool-dependent

The decision statistic we study is the maximum compatibility score over the retrieved candidates. A score-based policy retrieves a candidate set $\mathcal{C}_t$ and computes

$$m_t = \max_{c \in \mathcal{C}_t} s(x_t, c), \qquad \hat{c}_t = \arg\max_{c \in \mathcal{C}_t} s(x_t, c), \tag{1}$$

attaching x_t to $\hat{c}_t$ when $m_t \geq \tau$ and starting a new event otherwise. The maximum is an order statistic, so its value depends on how many candidates it is taken over and not only on the incoming document. To make this measurable, we use the first mention of a gold event and restrict attention to historical candidates annotated as different events. Let M_K be the maximum score over a ranked prefix of K such non-coreferent candidates, and define

$$\mu_K = \mathbb{E}[M_K], \qquad r_K(\tau) = \Pr(M_K \geq \tau). \tag{2}$$

Here $r_K(\tau)$ is a decision-level false-match risk. We do not claim that the compared documents are unrelated in every sense; we use the auditable benchmark fact that they are not coreferent. For an idealized pool of independent scores with CDF F, the maximum satisfies $\Pr(M_K \leq u) = F(u)^K$. We use this identity only as intuition, because real retrieval scores are dependent and query-specific; it explains why a threshold that is calibrated at one pool size loses meaning at another.

3.3 Two error types and the measures we report

Every streaming decision falls into one of two gold situations, and we separate them because our intervention affects them differently. A new-event decision should reject memory, and failing to do so is a false merge. A repeated-event decision should attach to an existing cluster, and failing to do so is a missed repeat. For a policy π we therefore report

$$\underbrace{\text{risk}(\pi) = \Pr[\pi(x_t) = \textsc{same} \mid x_t \text{ is new}]}_{\text{false merges}}, \quad \underbrace{\text{coverage}(\pi) = \Pr[\pi(x_t) = \textsc{same} \mid x_t \text{ is repeated}]}_{\text{repeated events recovered}}. \tag{3}$$

We additionally report cluster-level B-cubed F1 and the adjusted Rand index (ARI), which measure whether mentions are grouped into the right events overall.

3.4 Diagnostic protocol

The diagnostic is designed so that every compared pair is verifiably non-coreferent. For each query, historical mentions are ranked by the fixed retriever and the candidate pools $K \in \{4, 16, 64, 256\}$ are nested prefixes of that ranking. A query is the first mention of a gold chain, and its candidates are restricted to earlier mentions from other annotated chains and other source documents. An audit records the candidate chain identifiers and finds no same-chain violations. The monotonicity of a nested maximum is expected; the quantities of interest are how large the increase is and how much false-match risk it implies.

4 Method: Candidate-Conditioned Event Judge

Overview. The method keeps everything about the baseline except the final decision. For an arriving document we retrieve the same top-K source-excluded candidates and maintain the same memory, but instead of reducing the candidate scores to a single maximum and thresholding it, we hand the whole bounded candidate set to one judge and ask it to decide whether any candidate describes the same real-world occurrence. The judge answers either with the identifier of one candidate, in which case the document joins that candidate's cluster, or with UNSEEN, in which case it starts a new cluster. The document enters memory only after this decision, so the protocol stays online. Two design choices carry the argument: the candidate set is bounded and built by the same retriever as the baseline, so a change in results comes from the decision rather than from better retrieval; and the judge must return a supplied identifier or UNSEEN, so it cannot invent an event outside the retrieved pool. The rest of this section specifies candidate construction and the judge interface, and then explains why the design tests the proposed mechanism rather than merely swapping in a stronger model.

4.1 Candidate construction

Let e_t be the event-bearing mention and x_t its incoming report. We encode the report with a frozen BGE-M3 encoder and retrieve historical reports from the streaming memory. Candidates from the same source document are removed before ranking, because they are alternative descriptions within one article rather than independent cross-document evidence. The candidate set is

$$\mathcal{C}_t = \text{TopK}_{c \in \mathcal{H}_{t-1},\, \text{src}(c) \neq \text{src}(x_t)} \cos(v(x_t), v(c)), \qquad K = 8, \tag{4}$$

where $\mathcal{H}_{t-1}$ contains only reports observed before t, and the retriever is the same BGE-M3 retriever used by the baseline. The pool is not the bottleneck: on the three held-out WEC streams it contains at least one mention from the gold event for 86.0–87.3% of repeated-event decisions. This figure is computed for analysis only; gold chains never construct the pool.

4.2 Candidate-conditioned judgment

The judge makes a single bounded decision per document over the retrieved set. It receives the current report, its supplied event mention, and the text and identifier of each retrieved historical candidate, and returns one of two actions:

$$(a_t, \hat{c}_t) = J_\phi\left(x_t, e_t, \{(c, x_c, e_c) : c \in \mathcal{C}_t\}\right), \quad a_t \in \{\text{SAME}, \text{UNSEEN}\}. \tag{5}$$

The judge runs under a fixed instruction. It must resolve the incoming mention against the historical candidates and return JSON only, with `label`$\in$ {`SAME_EVENT`,`UNSEEN_EVENT`} and `target` set to a supplied candidate id or null. The instruction defines SAME as the same real-world occurrence rather than a related topic or event type, and warns against rejecting a genuine repeat merely because it is short, elliptical, or differently worded. The incoming report, its mention, and the candidate reports and identifiers are appended verbatim. Two constraints keep the interface well behaved: a SAME decision must copy a supplied identifier, and an invalid target is mapped to UNSEEN. The output space is therefore closed and the judge cannot invent a historical event.

The online update is therefore

$$z_t = \begin{cases} \text{cluster}(\hat{c}_t), & a_t = \text{SAME}, \\ \text{new_cluster}(), & a_t = \text{UNSEEN}, \end{cases} \qquad \mathcal{H}_t = \mathcal{H}_{t-1} \cup \{(x_t, z_t)\}. \tag{6}$$

The update rule keeps the method one-pass and prevents two ways of cheating. A decision is made before the current report enters memory; later retrieval uses only previously observed report text and source exclusion; and the predicted cluster label is not fed back into retrieval. Because the judge cannot attach to a candidate outside the retrieved pool, gains cannot come from an unbounded search over the benchmark.

4.3 Design rationale

The fixed BGE-M3 baseline reduces the candidate scores to one maximum and applies a global threshold. C2E-Judge keeps the same source-excluded retrieval and the same memory, but exposes the complete top-eight evidence set to a semantic judge operating on whole reports. This is a mechanism-consistent system test: if the failure is treating one accidental high score as sufficient evidence, a richer candidate-conditioned decision should reduce false merges without changing candidate generation. Because the two scorers use different text views, the result is not presented as a model-independent causal estimate; the held-out visibility and context controls separately probe the proposed explanation.

The contribution is the candidate-conditioned interface and the online protocol that makes the failure mode measurable. API dependence and token cost are reported explicitly rather than hidden behind the clustering scores.

4.4 Complexity and reproducibility

Retrieval costs one BGE-M3 encoding and a similarity search over the current memory. The judge makes at most one bounded call per incoming document and sees at most eight historical reports. In the full test runs, the method used 1,891 fresh API calls for 1,893 documents per stream, about 3.45–3.49 million input tokens, and about 35.6 thousand output tokens. The released prediction files, manifests, and per-order token counts permit metric reproduction without rerunning the API; rerunning the calls remains subject to endpoint-version changes. Of the two documents without a fresh request, one follows the deterministic empty-pool path and one is served by the released cache. An offline audit of all 5,673 recorded calls found zero API errors, invalid labels, or invalid same-event targets.

5 Experiments

5.1 Research questions and experimental setting

Our experiments answer three questions. **RQ1 (phenomenon):** does the maximum score over annotated non-coreferent candidates increase with candidate-pool size? **RQ2 (decision consequence):** does a fixed threshold acquire a different false-merge meaning as the pool grows, or is the drift only an uncalibrated score scale? **RQ3 (intervention):** with the retriever, memory, and candidate budget

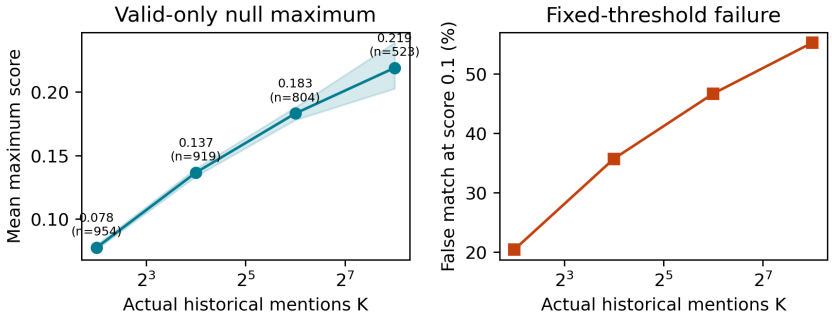


Figure 2: Valid-only candidate-pool false familiarity on three WEC test orders. All diagnostic candidates are non-coreferent and from different source documents. Increasing K raises the maximum score and makes a fixed threshold less selective.

fixed, does a judge that reads the whole candidate set improve over one that collapses the set to a maximum?

Benchmarks. The primary benchmark is WEC, a cross-document event-coreference collection derived from Wikipedia (Eirew et al., 2021). We project it into a streaming protocol: documents arrive one at a time and the memory contains only the observed prefix. We use the development split of order 17 for protocol and threshold choices and report on the disjoint test splits of orders 17, 29, and 43; the shared number is only a shuffle seed, and no test document is used for threshold or prompt selection. As a transfer check we use a binary ECB+ projection (Martinelli et al., 2025) under the identical protocol.

Baselines. We compare against the frozen official WEC pair scorer with source-excluded BGE-M3 retrieval and a development-selected threshold of 0.18. We additionally report score-only controls (whole-report cosine maxima, and a development-tuned pool-size threshold family) so that no gain is attributed to richer text or to threshold calibration alone.

Method under test. C2E-JUDGE keeps the same BGE-M3 encoder, source exclusion, streaming memory, and top-eight candidate budget, and sends the bounded candidate set to a frozen language-model judge. No benchmark label ever enters the prompt.

Metrics. We report B-cubed F1 and adjusted Rand index (ARI) for clustering, plus two operational measures: new-event false-merge risk and repeated-event coverage, defined in Section 3 (Eq. 3).

Hardware and cost. Retrieval and evaluation run on two NVIDIA RTX 4090 GPUs and an AMD Ryzen 9 5950X CPU; the judge is a remote frozen endpoint. The full three-order WEC runs use 5,673 API calls, roughly 10.4M input and 107K output tokens, and all statistics can be recomputed from released prediction files without new calls.

5.2 THE MAXIMUM RISES WITH THE CANDIDATE POOL (RQ1–RQ2)

Our first result is that the effect is real and monotone under a protocol in which every compared pair is verifiably non-coreferent. Figure 2 shows this. The protocol enforces source-document exclusion, and counting only queries with at least K actual candidates, the mean maximum non-coreferent score increases from 0.0776 at $K = 4$ to 0.2188 at $K = 256$. The fixed-threshold false-match rate rises from 20.4% to 55.2%. The valid query counts are 954, 919, 804, and 523; incomplete pools are not padded. The diagnostic threshold 0.1 is a deliberately fixed stress point for this curve, not the end-to-end baseline threshold 0.18. On the same 523 queries available at every pool size, the mean maximum rises from 0.0842 to 0.2190 and the false-match rate from 22.6% to 55.6%; the paired risk increase is 33.1 points (28.7, 34.9, and 35.6 points in orders 17, 29, and 43).

5.3 A CANDIDATE-CONDITIONED JUDGE IMPROVES THE STREAM (RQ3)

Table 1 holds the BGE-M3 retriever, source exclusion, memory, and top-eight budget fixed. The baseline uses the official WEC mention-centered scorer, whereas the proposed system interprets

Table 1: WEC test results under the online protocol, averaged over three stream orders. The baseline uses source-excluded BGE-M3 retrieval and the frozen official WEC pair scorer. C2E-JUDGE uses the same top-eight retrieval and a fixed `deepseek-flash` candidate judge.

Policy	B^3 F1 ↑	ARI ↑	Risk ↓	Coverage ↑
BGE-M3 + WEC max	0.5441	0.2501	16.8%	54.3%
C2E-JUDGE (proposed)	**0.7331**	**0.7077**	**0.4%**	**65.6%**
Absolute change	+0.1890	+0.4576	-16.4 pp	+11.3 pp

Table 2: Same-judge candidate-visibility dose response (mean over three orders). Same runner, prompt, retriever, and memory; only the visible candidate count K changes.

K	B^3 F1 ↑	ARI ↑	Risk ↓	Coverage ↑
1	.6020	.5143	0.0%	50.4%
2	.6261	.5648	0.0%	51.8%
4	.6952	.6626	0.2%	61.1%
8	**.7331**	**.7077**	0.4%	**65.6%**

whole-report candidate evidence with the language-model judge. This is an intentional system comparison; it does not isolate a model-independent effect of list length alone.

The proposed method improves both cluster metrics by more than ten percentage points, with a large reduction in false merges and higher repeated-event coverage; the retriever and top-eight budget are shared, while the judge sees the candidate reports jointly instead of treating one maximum as sufficient. A block jackknife gives the paired top-eight minus baseline gaps of $+18.9$ [17.8, 20.0] B^3 points and $+45.8$ [42.9, 48.6] ARI points, with risk and coverage gaps of -16.4 [-19.1, -13.6] and $+11.3$ [9.6, 13.0] points. For a direct same-judge control, we rerun the complete 1,893-document test streams with the same whole-report prompt and memory while varying how many retrieved candidates the judge can see (Table 2). Because candidate retrieval uses only observed report text and source exclusion, not previous predicted cluster labels, all four runs construct the same prefix candidate history; only candidate visibility changes. Cluster quality improves monotonically and saturates: the paired top-eight minus top-one gaps are $+13.1$ [12.0, 14.2] B^3 points and $+19.3$ [16.8, 21.9] ARI points, and each doubling step has a jackknife interval that excludes zero. False-merge risk stays at or below 0.4% at every visibility level while repeated-event coverage rises from 50.4% to 65.6%.

As score-only controls, whole-report cosine top-two aggregation (threshold 0.80) and a development-tuned rule $m_t \geq b + s\log(K/8)$, whose selected slope is $s = 0$ for every order, reach only B^3/ARI $(.5433, .3313)$ and $(.5545, .3490)$—both below the same judge's top-one $(.6020, .5143)$. A pool-size threshold correction therefore does not explain the gain. Decomposing the gain, the same-judge top-one to top-eight increase of 13.1 B^3 points accounts for 69% of the 18.9-point system-level change from Table 1. The visibility effect is therefore a substantial component of the reported improvement rather than a secondary artifact of the stronger decision model. The two controls fix different errors. On the 500-document prefixes top-one is already conservative (risk 0.0% at every level) while coverage rises from 46.1% to 57.5%, so visibility is a recall mechanism; the false-merge drop in Table 1 instead separates the judge from the score threshold. Both hold under a stricter prompt ($0.8237/0.6544$ B^3/ARI at top-eight vs. $0.8271/0.6570$ balanced) and a second judge endpoint ($+4.3$ B^3, $+9.9$ ARI, $+11.3$ coverage, risk 0.0% at both levels).

5.4 THE EFFECT TRANSFERS TO A SECOND DATASET

To test whether the intervention is specific to WEC, we apply the same source-excluded BGE-M3 stream protocol to a binary ECB+ projection, using a 500-document prefix from each of three fixed orders. With the threshold selected on the order-17 development stream, cosine reaches B^3/ARI $(.6861, .1570)$, while the same C2E interface reaches $(.8199, .3936)$; the gains are $+13.4$ and $+23.7$ points. This is a transfer check rather than a new leaderboard claim: the prefix is shorter, and the prompt endpoint is held fixed rather than tuned on ECB+.

5.5 Results are stable across orders, and retrieval is not the limit

The improvement is stable across the three fixed test orders: per-order B^3 spans 0.5326–0.5506 for the baseline and 0.7241–0.7423 for C2E-Judge, with no order inverting the gain. The top-eight retriever contains a gold-chain candidate for 87.3%, 86.7%, and 86.0% of repeated-event decisions in orders 17, 29, and 43, respectively. Thus the judge has a real retrieval ceiling, but the measured gain is not produced by giving it access to all historical documents.

As an uncertainty check, a delete-one jackknife over 57 contiguous 100-document blocks, rescoring recorded predictions without new API calls, gives C2E-Judge B^3 0.7332 [0.7223, 0.7440] and ARI 0.7077 [0.6817, 0.7338].

5.6 Controls confirm the mechanism

A paired analysis of 523 identical first-mention queries with at least 256 historical candidates finds a mean increase of 0.1410 in $M_{256} - M_4$ (median 0.0676), with a nondecreasing maximum for every query. Random subpools from the same retrieved candidates also show inflation, while the nested prefix is stronger at small K because retrieval rank matters. These controls support a narrow claim about candidate-pool decision behavior. They do not prove that every event-coreference system shares the same score scale.

To test whether the drift is merely an uncalibrated score scale, we replace the maximum with a pool-relative extremeness score $z_K = (m_K - \mu_{-1})/\sigma_{-1}$, where μ_{-1}, σ_{-1} are the mean and standard deviation of the query's remaining candidate scores. Calibrating both rules to the same $K{=}4$ risk, the raw maximum drifts from 22.8% to 55.6% between $K{=}4$ and $K{=}256$, whereas the pool-relative rule still drifts from 22.8% to 47.8%. Normalising against the query's own pool removes only 24% of the drift, so the failure is not a fixable threshold scale; most of it reflects genuine semantic proximity among non-coreferent candidates.

6 Conclusion and Limitations

We studied why a growing event memory makes new events look familiar, and what a system should do about it. On WEC, the maximum score over annotated non-coreferent candidates rises with the candidate-pool size, and a pool-relative normalisation removes only a quarter of that drift, so the problem is not a threshold that merely needs rescaling. We therefore tested C2E-Judge, which keeps retrieval and the top-eight budget fixed but lets a semantic judge read the whole candidate set. Across three held-out orders, B-cubed F1 rises from 0.5441 to 0.7331 and ARI from 0.2501 to 0.7077. Under one judge, top-one to top-eight visibility supplies 69% of that B^3 gain and raises repeated-event coverage from 46.1% to 57.5% at unchanged near-zero risk, and the visibility effect reproduces under a stricter prompt, a second judge endpoint, and a second dataset.

Our claims are bounded in three ways. First, the evidence covers supplied-mention WEC streams under one frozen retriever and one judge, so it does not extend to event detection or open-world retrieval. Second, the reported intervals come from block resampling of recorded predictions and do not capture endpoint drift. Third, the method costs one judge call per document. The takeaway is correspondingly narrow: candidate-conditioned comparison is a strong system-level intervention and a promising distillation target, not a new architecture or a universal solution.

AI-use statement: the language model is used only as the fixed experimental judge; all reported claims were checked by the authors.

References

Shafiuddin Rehan Ahmed, George Arthur Baker, Evi Judge, Michael Reagan, Kristin Wright-Bettner, Martha Palmer, and James H. Martin. Linear cross-document event coreference resolution with X-AMR. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10517–10529. ELRA and ICCL, 2024a. URL https://aclanthology.org/2024.lrec-main.920/.

Shafiuddin Rehan Ahmed, Zhiyong Eric Wang, George Arthur Baker, Kevin Stowe, and James H. Martin. Generating harder cross-document event coreference resolution datasets using metaphoric paraphrasing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 276–286, 2024b. doi: 10.18653/v1/2024.acl-short.27.

James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 37–45, 1998. doi: 10.1145/290941.290954.

Emily Allaway, Shuai Wang, and Miguel Ballesteros. Sequential cross-document coreference resolution. In *Proceedings of EMNLP*, pp. 4659–4671, 2021. doi: 10.18653/v1/2021.emnlp-main.382.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=33XGfHLtZg`.

Manish Baveja, Hongsong Yuan, and Lawrence M. Wein. Asymptotic biometric analysis for large gallery sizes. *IEEE Transactions on Information Forensics and Security*, 5(4):955–964, 2010. doi: 10.1109/TIFS.2010.2058105.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1563–1572, 2016. doi: 10.1109/CVPR.2016.173.

Chien Chin Chen, Meng Chang Chen, and Ming-Syan Chen. An adaptive threshold framework for event detection using hmm-based life profiles. *ACM Transactions on Information Systems*, 27(2):9:1–9:35, 2009. doi: 10.1145/1462198.1462201.

Xinyu Chen, Peifeng Li, and Qiaoming Zhu. Employing discourse coherence enhancement to improve cross-document event and entity coreference resolution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23272–23286, 2025. doi: 10.18653/v1/2025.acl-long.1134.

Xinyu Chen, Peifeng Li, and Qiaoming Zhu. Incorporating temporal coherence to cross-document event coreference resolution. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29151–29168, 2026. doi: 10.18653/v1/2026.acl-long.1345.

Alon Eirew, Arie Cattan, and Ido Dagan. WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. In *Proceedings of NAACL-HLT*, pp. 2498–2510, 2021. doi: 10.18653/v1/2021.naacl-main.198.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Conformal prediction sets with limited false positives. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 6534–6553, 2022.

Mohsen Jafarzadeh, Akshay Raj Dhamija, Steve Cruz, Chunchun Li, Touqeer Ahmad, and Terrance E. Boult. A review of open-world learning and steps toward open-world learning without labels. *arXiv preprint arXiv:2011.12906*, 2022. doi: 10.48550/arXiv.2011.12906.

Ying Jin and Emmanuel J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

Robert L. Logan IV, Andrew McCallum, Sameer Singh, and Dan Bikel. Benchmarking scalable methods for streaming cross document entity coreference. In *Proceedings of ACL-IJCNLP*, pp. 4717–4731, 2021. doi: 10.18653/v1/2021.acl-long.364.

Giuliano Martinelli, Bruno Gatti, and Roberto Navigli. xCoRe: Cross-context coreference resolution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34252–34266, 2025. URL `https://aclanthology.org/2025.emnlp-main.1737/`.

Qingkai Min, Qipeng Guo, Xiangkun Hu, Songfang Huang, Zheng Zhang, and Yue Zhang. Synergetic event understanding: A collaborative approach to cross-document event coreference resolution with large language models. In *Proceedings of ACL*, pp. 2985–3002, 2024. doi: 10.18653/v1/2024.acl-long.164.

Beatrix M. G. Nielsen and Lars Kai Hansen. Hubness reduction improves sentence-bert semantic spaces. In *Proceedings of the 5th Northern Lights Deep Learning Conference*, volume 233, pp. 181–204, 2024.

Xiaoning Wang, Yuyang Huo, Liuhua Peng, and Changliang Zou. Conformalized multiple testing after data-dependent selection. In *Advances in Neural Information Processing Systems*, 2024.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of EMNLP*, pp. 926–941, 2022. doi: 10.18653/v1/2022.emnlp-main.60.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. *arXiv preprint arXiv:2512.12844*, 2025. doi: 10.48550/arXiv.2512.12844.

Jin Zhao, Nianwen Xue, and Bonan Min. Cross-document event coreference resolution: Instruct humans or instruct GPT? In *Proceedings of the 27th Conference on Computational Natural Language Learning*, pp. 561–574, 2023. doi: 10.18653/v1/2023.conll-1.38.

Jin Zhao, Jingxuan Tu, Bingyang Ye, Xinrui Hu, Nianwen Xue, and James Pustejovsky. Beyond benchmarks: Building a richer cross-document event coreference dataset with decontextualization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3499–3513, 2025. doi: 10.18653/v1/2025.naacl-long.178.