

Turning message streams into trustworthy world models

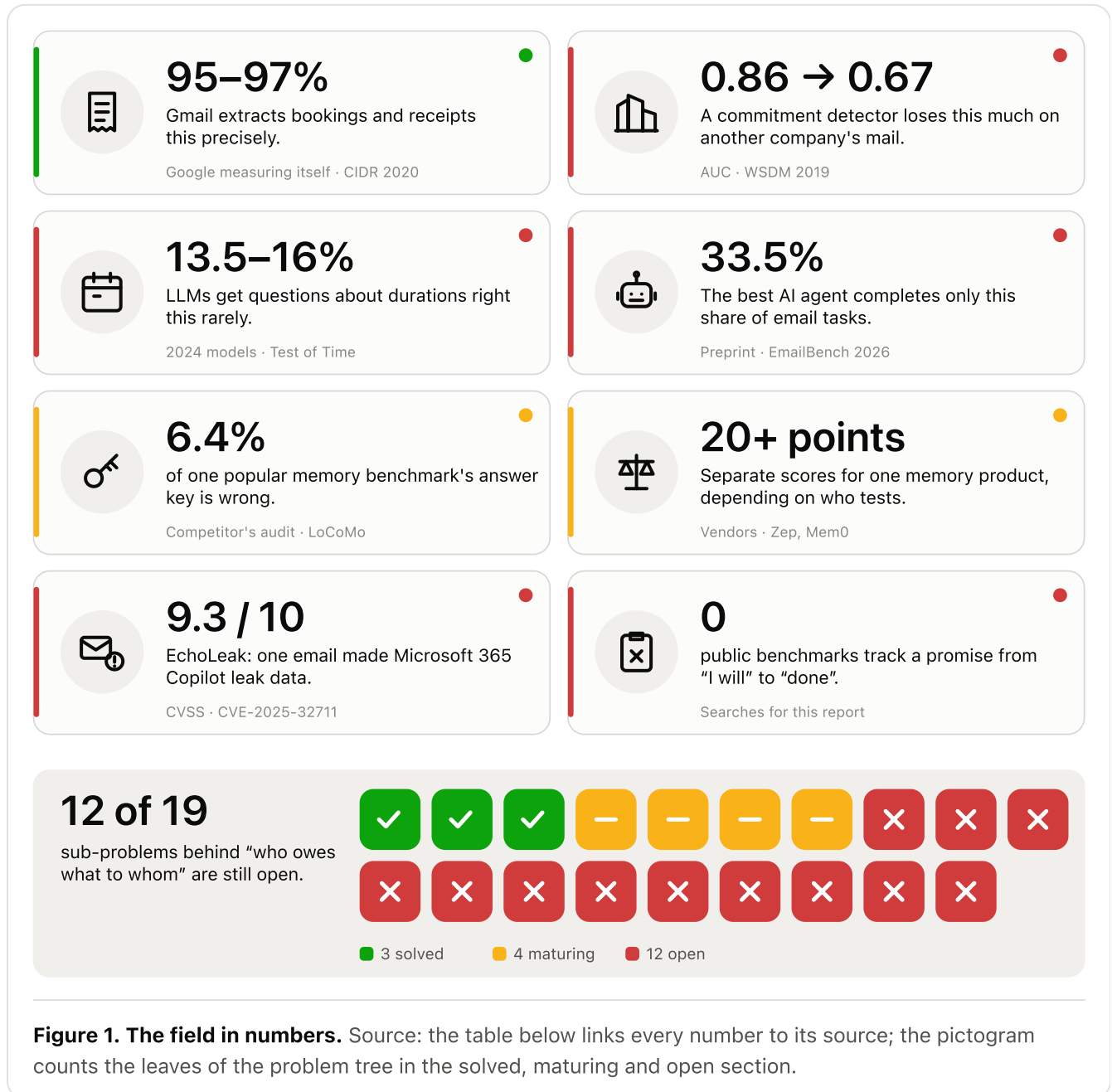
How people and machines turn chat, meetings and email into threads, topics, entities, events and commitments: the theory, the research, what shipped and what retreated, what is solved and what is still open, and a reference architecture for building it.

Extracting a "world model" from chat, meetings and email is half solved: the structural plumbing is industry standard, while the semantic layer people actually want stays open.

In brief

- **Commodity:** header-based [threading](#), [quoted-text deduplication](#), templated extraction from machine-generated email ([G](#) Google's own engineers report 95-97% [precision](#), vendor-authored: [Sheng et al., CIDR 2020, Table 2](#)), and meeting transcripts with summaries.
- **Not solved: human-to-human commitments.** That means detecting who promised what to whom, by when, whether it was accepted and whether it was ever done.
 - No peer-reviewed [LLM-era benchmark](#) measures commitment or action-item extraction from real email or meetings.
 - The best pre-LLM commitment classifiers lose much of their accuracy when moved from one company's mail to another's.
- **The theory has been around for decades.** [Speech-act theory](#), [Clark's grounding](#), [Singh's social commitments](#), [bi-temporal databases](#) and [truth maintenance](#) give a precise vocabulary for what to extract and how to store it.
- **Engineering lesson:** a world model built from messages has to be a **revisable, provenance-linked projection over an immutable log, not a set of facts.**
 - Every derived claim should point back to its source message and carry its confidence and extractor version.
 - Claims are invalidated rather than deleted, and a human should be able to contest them.
- **Three further constraints shape any design:**
 - Ingested email is a proven attack channel ([EchoLeak](#), [CVSS 9.3](#)).
 - Published memory-benchmark numbers for the same product vary by 20 points or more depending on who ran them.
 - No regulator has yet said how [GDPR erasure](#) reaches derived graphs and summaries.

Key numbers



Number	What it measures	Source
95-97%	Precision of G Google's extraction from machine-generated email (events 95.3, purchases 97.5), measured by Google on its own system	Sheng et al., CIDR 2020, Table 2
AUC 0.86 → 0.67	A commitment classifier moved from one company's mail (Avocado) to another's (Enron)	WSDM 2019
kappa ~0.7	Human agreement on email requests and commitments, which caps achievable accuracy	NAACL 2010 ; arXiv 2005.06282
0	Peer-reviewed 2023-2026 benchmarks for LLM commitment or request detection in email	Two targeted searches (see below)

Number	What it measures	Source
33.5% of 206	Scenarios passed by the best EmailBench agent configuration (preprint seen via search)	✗ arXiv 2609.31906
13.5-16%	Frontier-model accuracy on duration arithmetic (synthetic benchmark)	ICLR 2025
20+ points	Spread of memory-benchmark scores for the same product, depending on who ran the test	Zep blog ; Mem0
CVSS 9.3	Severity of EchoLeak, a zero-click attack via one crafted email	The Hacker News

Scope and evidence note

- This report synthesises eight research notes compiled in October 2026.
- "Vendor-reported" figures were produced by a company about its own product (or, where stated, a competitor's).
- Items the notes could not verify are flagged or left out. Several 2026 [✗ arXiv](#) preprints were seen only through search summaries; they are marked where cited.

CONTENTS

Summary: main conclusions and limits	Ingested messages are an attack surface and a liability
Speech acts and grounding define what is worth extracting	Things you probably have not thought of
Machine pipelines work stage by stage, with uneven maturity	Navigating the field
Threading is a workflow, not a step	Conclusion
What shipped, what retreated, and how practitioners work	What works and what does not
The solved, maturing and open map	Mind map of the report
A reference architecture: log, claims, beliefs, review	About this report and its session log
	Glossary

Summary: main conclusions and limits

Limits of this research

- **Some sources were read as search snippets, not opened.** Several 2026 preprints are unverified; the report flags each one where it is cited.
- **There is no published test of current [LLMs](#) on commitments or meeting action items.** No peer-reviewed benchmark measures how well today's models pull commitments out of real email, or action items out of real meetings. The report says so rather than guessing a number.

- **Vendor numbers are labelled.** Where a company measured its own product, or a competitor's, the report says so.

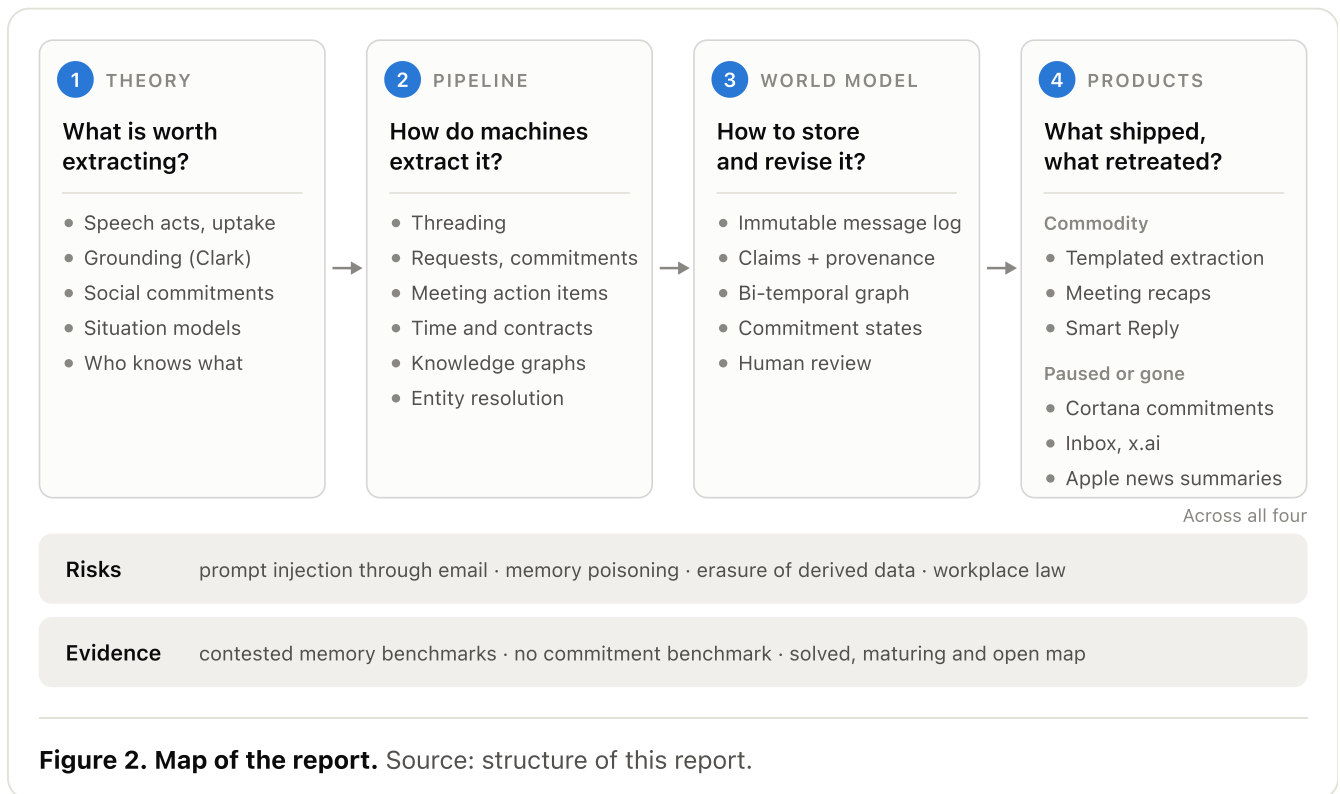
Main conclusions

- **Solved and industry standard:**
 - rebuilding email threads from headers
 - spotting the "inclusive" email that already quotes the whole thread, a legal-discovery technique
 - extracting data from machine-generated email such as bookings and receipts ([G](#) Google's own engineers report 95-97% precision, vendor-authored: [Sheng et al., CIDR 2020, Table 2](#))
 - meeting transcripts and summaries, and action items that people state explicitly
- **Open:** tracking human commitments through their whole life: who promised what, to whom, by when, whether it was accepted, and whether it was done.
 - Pre-LLM classifiers lose much of their accuracy when moved to another company's email: AUC falls from 0.86 to 0.67 ([Azarbonyad, Sim & White, WSDM 2019](#)).
 -  Microsoft paused its commitments email (Briefing) on 15 January 2023 and gave no reason ([Microsoft support](#)).
 - Resolving relative dates like "next Friday" and reasoning about durations are weak points for LLMs: 2024 frontier models answered duration questions correctly only 13.5-16% of the time ([Test of Time, ICLR 2025](#)).
- **The theory is ready to use:**
 - [Speech-act theory](#): a request is not a commitment until someone accepts it.
 - [Clark's grounding](#): shared understanding needs evidence that it was understood.
 - [Singh's commitment state machines](#) give a ready schema for commitments.
 - Situation models suggest the dimensions to track: time, place, entity, cause and intent.
 - [Transactive memory](#) supports a "who knows what" index.
 - Suchman's critique explains why such systems feel like surveillance.
- **Architecture (synthesis, labelled as such below):**
 - Keep an [immutable message log](#).
 - Store extracted claims separately, each with its source and confidence.
 - Build the current world state from those claims, keeping a history of when each fact was true and when it was recorded; old facts are marked invalid rather than deleted.
 - Merge duplicate people and organisations in a way that can be undone, and have humans review the uncertain cases.
- **Benchmarks for memory products are unreliable.** Scores for the same product differ by 20 points or more depending on who ran the test ([Zep](#), [Mem0](#), both vendors), and a competitor's audit found 6.4% of one popular benchmark's answer key wrong ([Penfield Labs](#)). Only trust tests rerun on your own data.
- **Risks:**
 - Every incoming email is a way to inject instructions into the system. EchoLeak (CVE-2025-32711, [CVSS 9.3](#)), a zero-click attack on  Microsoft 365 Copilot through one crafted email, is the proven case ([The Hacker News](#)).
 - Planted false memories can persist and affect later answers ([MINJA](#), [X](#) [arXiv 2503.03704](#); [SpAIware](#)).

- No regulator has said how the [GDPR right to erasure](#) applies to summaries and graphs derived from messages; the closest guidance is [EDPB Opinion 28/2024 on AI models](#) ([CMS summary](#)).

How the report is organised

The three figures below give the overview. The first maps the report: the four questions it answers in order, and the two concerns that cut across all of them. The second places the main papers, products and incidents on one timeline. The third takes eight everyday problems and sets how people cope next to how machines try, and where the machines are known to fail. A mind map of the whole report closes it.



● Theory and human studies ● Research and benchmarks ● Products and incidents

Foundations, 1962–1999

- 1962 ● **Austin** shows that saying something can be doing something, like promising
- 1980s ● **The Coordinator** makes users tag each message as a request, promise or report
- 1991 ● **Clark & Brennan** explain how people confirm they were understood (grounding)
- 1991 ● **Wegner** shows that groups remember who knows what (transactive memory)
- 1993 ● **Suchman** warns that speech-act categories become a tool of control
- 1996 ● **Whittaker & Sidner** find that people use the inbox as a to-do list
- 1998 ● **Zwaan & Radvansky** show that readers track time, place, people, cause and intent
- 1999 ● **Singh** defines a commitment as a debt one party owes another

Classic NLP and first products, 2004–2021

- 2004 ● **Cohen, Carvalho & Mitchell** classify emails as requests, promises and other acts
- 2010 ● **Lampert, Dale & Paris** lift request accuracy from 72% to 84% by removing quoted text
- 2013 ● **Gmail** starts pulling flight and order details out of booking emails
- 2016 ● **Smart Reply** suggestions are used in 10% of replies sent from Inbox on mobile
- 2017 ● **Cortana** starts reminding users of promises found in their sent mail
- 2019 ● **Kummerfeld et al.** publish 77,563 hand-linked chat messages for disentanglement
- 2019 ● **Azarbonyad, Sim & White** show commitment detectors lose accuracy on another firm's mail
- 2019 ● **Google** shuts down Inbox by Gmail, its experimental email app
- 2020 ● **Google** reports that machine-learned extraction reaches 95–97% precision
- 2021 ● **Bizzabo** acquires x.ai, maker of the Amy scheduling assistant

LLM era, 2023–2026

- 2023 ● **Microsoft** pauses the Briefing email that listed users' commitments
- 2023 ● **Greshake et al.** show that hidden instructions in content can hijack LLM apps
- 2025 ● **Apple** pauses news notification summaries after they invent headlines
- 2025 ● **Mem0** scores Zep at 66.0 on LoCoMo; Zep's own rerun gives 75.1
- 2025 ● **GraphRAG-Bench** finds that graphs help with reasoning but not with simple lookups
- 2025 ● **EchoLeak** shows one crafted email can make Microsoft 365 Copilot leak data
- 2026 ● **Gmail's AI Inbox** starts suggesting to-dos, first for US English users
- 2026 ● **EmailBench**: the best agent completes only 33.5% of email tasks

Figure 3. Timeline of the field. Source: dates and figures as cited in the sections below.

Problem	 How people cope	 How machines try	 Known failure mode
Keeping promises	Rely on memory and flags; promises live in sent mail	Detect commitments in sent mail (Cortana, 2017)	Accuracy drops on another company's mail (AUC 0.86 → 0.67); Microsoft paused the Briefing email in 2023
Reading indirect requests	Infer intent from politeness, hedges and context	Classifiers and LLMs label the intent of each message	Hedged and conditional requests split even human annotators (kappa ~0.68)
Knowing it was accepted	Grounding: look for evidence the other side understood	Treat each detected request as a task	Uptake is rarely written down, and no benchmark measures it
Anchoring time	Read "next Friday" against when it was said	LLMs resolve dates from the text	LLMs get durations right only 13.5–16% of the time; no benchmark uses message timestamps
Finding things again	Scroll and search; heavy folder users do no better	Thread, search and auto-sort mail into bundles (Inbox)	Automatic sorting took control away from users; Inbox closed in 2019
Knowing who knows what	Transactive memory: ask the right person	Infer roles and experts from who writes to whom	Hierarchy inference was tested on one company (Enron) only
Summarising	Write meeting notes and recap emails	LLMs write recaps and notification summaries	Apple's summaries produced false headlines and were paused; GPT-4 Turbo spotted hallucinations with only 72% accuracy
Labelling the conversation	Leave acts implicit; people resisted declaring them (The Coordinator)	Infer acts automatically	It feels like surveillance (Suchman's critique, the Productivity Score backlash)

Figure 4. How people cope, how machines try, and how they fail. Source: the sections of this report on human practice, machine pipelines and industry; every entry is cited there.

Speech acts and grounding define what is worth extracting

The oldest and still most useful idea is that a message is an action, not just content, and a request is not a commitment until there is evidence it was accepted.

What a message does. Together these set the problem's ceiling for a builder.

Theory	Core idea	What it means for a machine	Source
Austin: speech acts	What an utterance says differs from what it <i>does</i> (asserting, promising, ordering)	A message is an action, not just content	SEP Speech Acts
Austin: felicity conditions	A request or promise is fully constituted only through uptake : the addressee has to take it up, or the act "misfires"	A request is not a commitment until there is evidence it was accepted	SEP Speech Acts
Indirect speech acts	"Could you maybe look at..." depends on inferring intention; some theorists treat it as implicature rather than a fixed form	Surface patterns will miss polite, hedged and conditional requests	SEP
Grice's maxims	Explain why meaning so often exceeds the literal text	Same ceiling: the literal text is not the whole request	Stanford handout

The Coordinator, and Suchman's critique

- Winograd and Flores turned speech acts into a product, The Coordinator.
 - Users labelled each outgoing message with a conversation category; the system tracked pending commitments and project status ([TOIS 1988 index](#), secondary description).
 - Its successor, ActionWorkflow, modelled work as loops of request, commit, perform, declare complete and accept ([Information Society 1993](#)).
- Lucy Suchman's critique is the founding warning for this whole field.
 - Building systems on speech-act categories brings "**an agenda of discipline and control**" over the people categorised.
 - The real politics lie in who defines the categories ([EUSSET abstract](#)).
- Winograd replied that explicit structure, like accounting procedures, supplies uniformity where ambiguity cannot be settled face to face ([Winograd 1994, search summary](#)).
- A 2010 essay says the Coordinator's demand for accountable communication limited it to "fairly small and dedicated networks" (💎 [ACM Interactions, snippet only](#)).
- No adoption figures survive.
- The durable lesson: **infer acts from natural messages instead of asking users to declare them, while keeping the inference contestable.**
- Automation removes the labelling burden but keeps the surveillance concern.

Commitments as data. Multi-agent systems research gave commitments a data model. These are ready-made schemas, not just philosophy.

Theory	Core idea	What it means for a machine	Source
Singh: social commitments	Mentalistic (belief and intention) semantics fail in open systems; commitments are directed (a debtor owes a creditor), arise in a context and can be manipulated in standard ways	A commitment record has a debtor, a creditor and a context	Singh IJCAI-99; Singh AAI-08
Castelfranchi vs Singh on acceptance	Castelfranchi required explicit acceptance by the creditor; Singh considers that right only for some applications	Whether acceptance is required depends on the application	Commitments in MAS
Yolum and Singh: commitment machines	Compile commitment-based protocols into finite-state machines	Allows exceptional moves that fixed message sequences forbid	Singh group
Symboleo	Same approach for legal contracts: state machines for the lifetimes of contracts, obligations and powers	Handles subcontracting, assignment and novation	Symboleo RE'20

- **Gap:** contract-law tests (offer, acceptance, [consideration](#)) were not sourced in the notes.
 - So the line between a social promise and a legally binding obligation remains a gap for human or legal review.

Grounding and discourse. Clark and Brennan add the epistemic side; [discourse theories](#) supply structure.

Theory	Core idea	What it means for a machine	Source
Clark and Brennan: grounding	Shared understanding is built step by step; each contribution needs evidence that it was understood "to a level sufficient for current purposes"; the medium changes how costly that evidence is	Email is asynchronous, so acknowledgement often never arrives. One message does not establish common ground	Clark & Brennan 1991
Grosz and Sidner	Intentional and attentional structure, including a focus stack	With RST and SDRT : a new message usually attaches to a recent open topic, a useful prior when linking messages	Grosz-Sidner discussion
RST and SDRT	Rhetorical relations	As above	SDRT intro

Humans keep the world model in memory, inbox position and other people

Empirical HCI work shows how people actually cope, and it argues against designs that need users to file things.

Study	What it found	Design implication	Source
Whittaker and Sidner (CHI '96)	Email doubles as a to-do list and archive; filing failures cause messages to be "lost"	Argues against designs that need users to file things	CHI '96
Ten-year follow-up, 600 mailboxes	Inbox sizes roughly unchanged while archives grew about tenfold	As above	 Microsoft Research
Field study, 345 users, more than 85,000 refinding actions	Heavy folder users were no more successful at refinding and were marginally slower; search and threading helped; scrolling was the most common behaviour	Search and threading help; folders do not	bespacific summary ; abstract-level only
Bellotti's Taskmaster (CHI 2003)	Grouped email, links and attachments into "thrasks" (thread plus task); users received it well	It never became a full product	CHI 2003
 Microsoft Research on commitments	Promises live in <i>outgoing</i> mail; people "just rely on memory" or on flags to track them	An open-commitments ledger replaces a fragile human function, and is useful only if it can tell an open commitment from a discharged one	MSR blog

Cognitive science suggests what dimensions a world model should track.

Theory	Core idea	What it means for a machine	Source
Zwaan and Radvansky: event indexing	Readers build situation models , not textbases , indexed by time, space, entity, causation and intentionality ; long-term memory is a network of event nodes linked by shared indexes	These are the dimensions to track	Zwaan & Radvansky 1998
Wegner: transactive memory	Groups spread memory across members through a directory of "who knows what"; couples who divided memory work naturally did better than strangers, and imposed assignments hurt performance	Supports a "who knows what" index	Wegner, Erber & Raymond 1991
Pirolli-Card sensemaking	Separates a foraging loop from a schema-building loop	A machine world model should behave like a set of hypotheses revised as evidence arrives, not like a ledger of settled facts	w Wikipedia summary

For a builder: do not make users file things; an open-commitments ledger must tell open from discharged commitments, and its contents are hypotheses revised as evidence arrives, not settled facts.

Machine pipelines work stage by stage, with uneven maturity

Each stage works where structure or schema is tight and breaks on open extraction, time and commitments.

Stage	Where it works	Where it breaks
Threading	Email with Message-ID, In-Reply-To and References headers	Stripped headers; chat without reply structure; topics that span threads
Email speech acts and commitments	Within one corpus	Across organisations; no LLM-era benchmark
Meetings	Summaries are measurable	Action items: little labelled data, no verified LLM score
Dialogue IE, time and contracts	Fixed schemas; formal contracts	Duration arithmetic and relative dates; obligations in informal dialogue
Graphs, entity resolution, memory	Classical record matching; graphs on multi-hop and summary questions	No merge-error rate on conversational data; long context degrades

Threading is deterministic when headers survive and hard when they do not

Email threads rebuild deterministically from headers; chat without reply structure, and threading by content alone, remain hard.

- Email threads are best rebuilt from Message-ID, In-Reply-To and References headers, with content similarity as a fallback.
- Some clients strip those headers, and threading on subject alone merges unrelated conversations ([mail-threading docs](#)).
- Chat without reply structure needs [disentanglement](#), framed as predicting reply-to links and then clustering.
- The reference benchmark is Kummerfeld et al.'s [77,563 annotated Ubuntu IRC messages](#) (✗ [arXiv 1810.11118](#)).
 - That work also found that about 80% of conversations in the widely used Ubuntu dialogue corpus were missing messages or contained extra ones.
- Getting whole threads exactly right is still hard:

Method or system	Year	Result	Caveat	Source
Supervised models, Kummerfeld test set	—	Cluster F1 36.6-51.7	—	—
Prompted GPT-4.1 and Gemini-2.5-pro	2026	Cluster F1 47.5 and 61.3	Workshop paper; not independently replicated	CEUR Vol-4178
Out-of-the-box GPT-4o	2026	"Performs poorly"	LREC 2026 study	ACL Anthology 2026.lrec-1.229

Method or system	Year	Result	Caveat	Source
Fine-tuned GPT-4o Mini	2026	74.4 link F1, but only 45.3 cluster F1	Same LREC 2026 study	ACL Anthology 2026.lrec-1.229
Zero-shot ChatGPT and GPT-4	—	Trailed supervised BERT by 25.6 and 12.3 points	Task is IRC speaker identification	✗ arXiv 2310.16301

- **Threads are not topics.** A 2026 study of an LLM assistant used by 105 network engineers found that "a single thread can span multiple unrelated topics"; GPT-4.1 split its 815 threads into 1,267 topic segments, with no accuracy check of that split reported ([Nakamura & Eguchi, NOMS 2026](#), ✗ [arXiv 2605.19627](#)). *Correction in v0.1: v0.0 called this a Slack case study; it is not about Slack.*
- The notes found no benchmark for linking one topic across threads.

For a builder: use platform-native thread IDs as primary keys for *structure*, and treat any content-inferred thread as a revisable cluster.

Email speech acts and commitments: mature methods, low ceilings, poor transfer

Classifying email by intent is a 20-year-old, small-data field: [human agreement caps accuracy](#), and models lose accuracy across organisations.

Method or system	Year	Result	Caveat	Source
Cohen, Carvalho and Mitchell: verb-noun taxonomy (Request, Propose, Commit, Deliver × Meeting, Information)	2004	Precision above 80% and recall above 50% for some classes	In the full paper, Commitment F1 ranged only 0.11-0.47; annotator kappa 0.72-0.83	EMNLP 2004 ; PDF
Lampert, Dale and Paris: request classifier on Enron	2010	72.3% → 83.8% accuracy once email "zoning" separated quoted, forwarded and signature text (a 41%+ error reduction)	Human agreement only kappa 0.681; conditional or indirect requests drove the disagreement	NAACL 2010
Azarbonyad, Sim and White ( Microsoft): sentence-level commitments	2019	F1 0.78-0.81 within one corpus	Avocado → Enron: F1 fell to 0.73 and AUC fell from 0.86 to 0.67 ; adversarial domain adaptation recovered part of the loss	WSDM 2019
Smart To-Do: commitment sentences on Avocado	—	86% precision, 84% recall	Annotator kappa 0.694	✗ arXiv 2005.06282
EmailBench : eight agent configurations	2026 (September)	Best configuration passed 33.5% of 206 scenarios	99.7% of its tool calls succeeded; preprint seen via search	✗ arXiv 2609.31906

- Lampert, Dale and Paris showed that preprocessing matters more than the model.
- Three conclusions follow:

- Human agreement near kappa 0.7 caps achievable accuracy.
- Zoning is a prerequisite.
- Per-organisation evaluation is unavoidable.
- Two targeted searches found **no peer-reviewed 2023-2026 benchmark for LLM commitment or request detection in email**. The closest is EmailBench.

For a builder: zone quoted, forwarded and signature text first, expect a ceiling near human agreement (kappa about 0.7), and evaluate per organisation.

Meetings: summaries are measurable, action items barely are

Summary errors are measured; action-item detection has very little labelled data and no verified LLM score.

- AMI has no direct action-item labels. Its 381 items across 101 meetings are derived by linking dialogue acts to summary sentences ([Liu et al., ICASSP 2023](#)).
- ICSI's annotations are no longer public (same source).
- Purver and colleagues found flat labels performed poorly and replaced them with **utterance roles: task description, timeframe, owner and agreement** ([SRI, MLMI 2006](#); [SIGDIAL 2007](#)).
 - That scheme remains the closest annotated model of an informal commitment.
- LLM-era work concentrates on summary errors:

Method or system	Year	Result	Caveat	Source
GPT-4 Turbo on the QMSum Mistake set (200 LLM summaries, 9 error types)	—	About 89% average accuracy detecting errors	Only 72% for hallucination	 arXiv 2407.11919
FRAME (structured pipeline)	—	Cut hallucination and omission scores by 2 of 5 points	—	 arXiv 2509.15901
Meeting Delegate, GPT-4o	—	Phonetically corrupted names, a stand-in for ASR errors, cut the explicit-cue response rate from 94.3% to 68%	Matters for owner extraction, but the paper did not measure action items, so the effect on assignee extraction is inferred	 arXiv 2502.04376

- The notes found **no verified LLM action-item F1 on AMI or ICSI**.

For a builder: Purver's utterance roles (task description, timeframe, owner, agreement) are the closest annotated model of an informal commitment; ASR name errors likely hurt owner extraction (inferred), and no published action-item score exists.

Dialogue IE, time and contracts: strong where the schema is tight

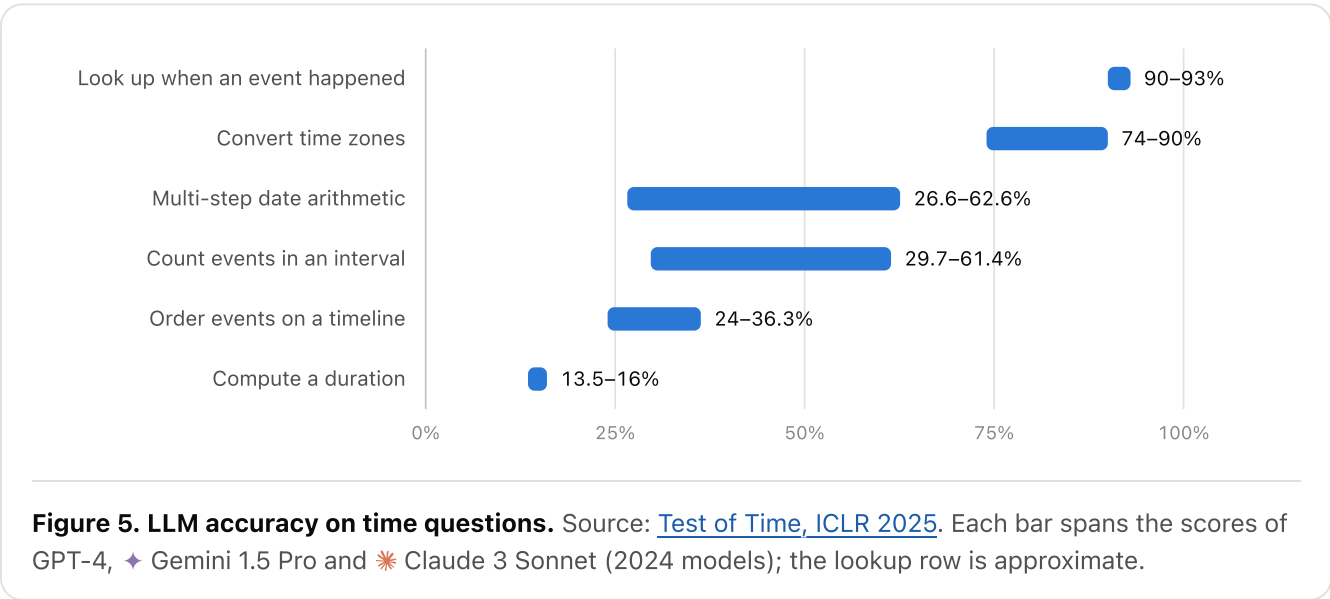
Where the schema is fixed, conversation-to-state is close to solved; open relations, time and informal obligations are not.

Method or system	Year	Result	Caveat	Source
MultiWOZ 2.4 (cleaned labels)	—	Raised every state-of-the-art tracker's score	—	arXiv 2104.00773
Qian et al.: MultiWOZ label fixes	2021	Correcting inconsistent tags added 7-10% joint goal accuracy	Unseen entities cut it by 29%, a sign of entity memorisation	Qian et al., SIGDIAL 2021
BERT on DialogRE (open relation extraction from dialogue)	—	61.2 F1 overall	55.4 when relations had to be found early in a conversation	Papers with Code
LLMs on DialogRE	2024	Competitive	Weak on inverse relations	IJCAI 2024

Time is a specific weak spot. In the synthetic [Test of Time benchmark \(ICLR 2025\)](#), frontier models scored:

Task	Score
"When did X happen" lookups	About 90-93%
Timelines	24-36%
Duration arithmetic	13.5-16% , with frequent off-by-one-day errors

- The paper's text and table disagree on one GPT-4 number (20.33% vs 40.25%).
- Relative expressions are hard for every system ([Frontiers](#)).
- The notes found **no benchmark for resolving "next Friday" against a message timestamp**.



Contracts. Contract clause extraction from formal documents works at "junior legal assistant" level.

Method or system	Year	Result	Caveat	Source
ContractEval: 19 LLMs on CUAD (510 contracts, 41 clause types)	—	Open models over-answered "no related clause"	Thinking mode lowered correctness	X arXiv 2508.03080
Another CUAD study, best configuration	—	Recall 0.77, precision 0.66	—	X arXiv 2505.12837
CraigslistBargain	—	Nearest relative for obligations in informal dialogue	Captures only the agreed price	X arXiv 1808.09637

- The notes found no benchmark for obligation extraction from informal dialogue.

For a builder: the design inference is to let the LLM output the unresolved time expression and to do the date arithmetic deterministically.

Graph construction, entity resolution and memory systems

Graph construction, entity resolution and [agent memory](#) work, but none is measured on conversational data, and graphs earn their cost only on multi-hop and summary questions.

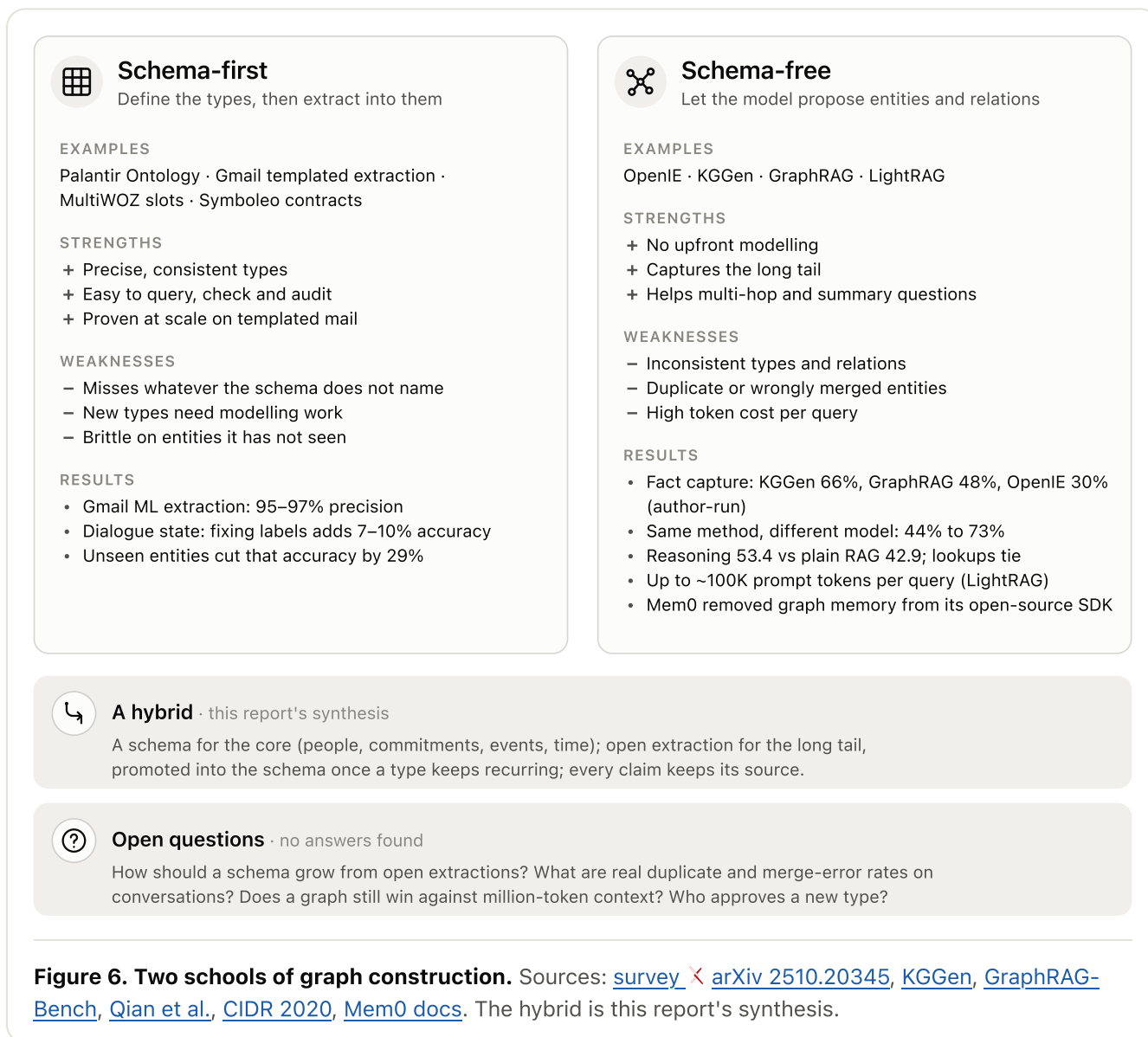
LLM [knowledge-graph](#) construction comes in two schools ([survey](#) [X arXiv 2510.20345](#)):

School	Strength	Weakness
Schema-based extraction	Precise	Scales poorly
Schema-free extraction	Flexible	Inconsistent

Method or system	Year	Result	Caveat	Source
KGGen, on its author-run MINE benchmark	—	Fact capture 66% for KGGen, 48% for GraphRAG and 30% for OpenIE ; backbone models range from 44% to 73%	Authors acknowledge that "over or under de-duplication" remains; test texts are encyclopedic articles, not conversations	X arXiv 2502.09956
Classical Fellegi-Sunter matching (Splink)	—	Mature for person and organisation identity; trains unsupervised and scales past 100M records	—	PyPI
Generative LLMs for entity matching (Peeters et al.)	2025	Match or beat fine-tuned models with little training data; generalise better to unseen entities	—	Peeters et al., EDBT 2025

Method or system	Year	Result	Caveat	Source
Senzing "principle-based" resolution	—	Reasons about how attributes behave, for example shared values versus stable ones	Vendor claims	Senzing

- The notes found **no measured duplicate or merge-error rate on conversational data.**



Agent-memory products fall into three families.

Family	Systems
Graph-indexed RAG	 GraphRAG, LightRAG, HippoRAG 2
Conversational fact stores	Mem0, Zep/Graphiti, MIRIX

Family	Systems
Agent-managed files	Letta,  Anthropic's memory tool

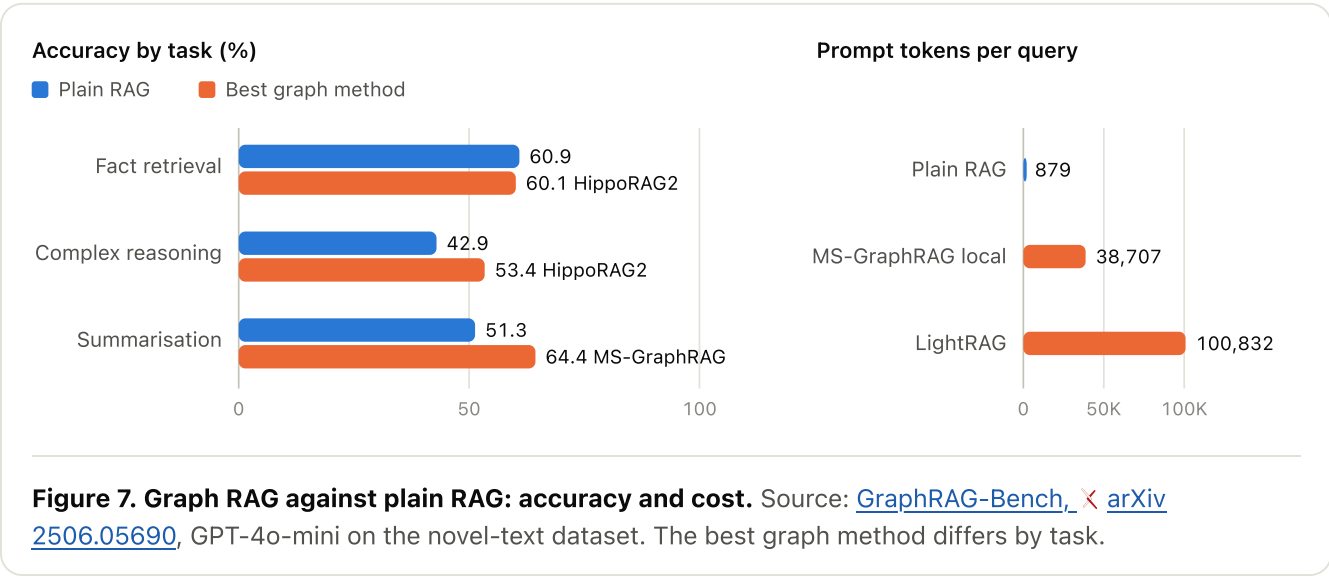
- Of the systems verified, **only Graphiti has first-class bi-temporal edges that invalidate old facts rather than deleting them** ([Graphiti](#)  [GitHub](#)).
- Graphiti also needs strong structured-output models; smaller ones often fail the schema.
- Mem0 has since **removed graph memory from its open-source SDK**, keeping entity linking only as a retrieval signal ([Mem0 migration docs](#)).

Graphs earn their cost on multi-hop and summary questions, not lookups. [GraphRAG-Bench](#) ( [arXiv 2506.05690](#)):

Task	Plain RAG	Best graph method
Fact retrieval	60.9	HippoRAG2 60.1
Complex reasoning	42.9	HippoRAG2 53.4
Summarisation	51.3	 MS-GraphRAG 64.4
Prompt tokens per query	About 880 (vanilla RAG)	38,700 ( MS-GraphRAG local search); 100,800 (LightRAG)

Long context is not a free alternative.

- NoLiMa found 11 of 13 models fell below half their short-context score at 32K tokens, with GPT-4o dropping from 99.3% to 69.7% ( [arXiv 2502.05167](#)).
- Accuracy is also lowest for information in the middle of the context ( [arXiv 2307.03172](#)).
- Those results come from 2024-25 models.



For a builder: use a graph for multi-hop and summary questions, not lookups, and budget for its token cost; long context is not a free alternative, and no merge-error rate on conversational data exists yet.

Threading is a workflow, not a step

Threading is a chain of six decisions, from "which text in this email is new" to "is this the same matter as last month's", and only the first two are solved.

- Each stage answers a different question from different signals, and its mistakes flow into every later stage.
- The TDT programme measured that cascade: automatic story boundaries raised topic-detection cost by 97% on broadcast audio, compared with boundaries given by hand ([Fiscus et al. 1999](#)).
- Joty et al. showed that a topic segmenter reading email in time order fails when topics interleave, and that giving it the reply structure from the threading stage improves it significantly ([Joty, Carenini & Ng, JAIR 2013](#)).
- This chapter expands the threading subsection above. It takes the approaches family by family (email, chat, then topics and stories), says which worked and which failed, and ends with what was disproven, a timeline and advice.

The workflow

Stage	What it decides	Input signal	Main method families	Maturity
1. Normalise and deduplicate	It decides which text in a message is new, which is quoted, and which copies are the same message.	Message bodies, quoted blocks, signatures and near-identical copies carry the signal.	Quote detection, email zoning and near-duplicate matching split and compare the text (Repke & Krestel 2018 ; Relativity).	It is solved in eDiscovery products, though footers, Bates stamps and OCR errors still create phantom messages (Relativity).
2. Link replies from headers	It decides which earlier message each message answers.	Message-ID, In-Reply-To and References headers,  Microsoft's Thread-Index and platform reply IDs name the parent.	The jwz algorithm and IMAP THREAD (RFC 5256) build a reply tree from those headers (RFC 5256).	It is solved while headers survive; some mail servers strip them, and a mailing list cannot repair them afterwards (Mailman wiki).
3. Reconstruct links from content	It decides which thread a message without usable headers belongs to.	Subject, participants, time gap, quoted text and body words carry the signal.	IR text matching, weighted similarity clustering and pairwise classifiers score candidate threads.	It is maturing in eDiscovery tools, which combine headers with body text, but no public benchmark exists and Enron has no true thread labels (Jamison & Gurevych 2013).

Stage	What it decides	Input signal	Main method families	Maturity
4. Untangle interleaved chat	It decides which conversation each chat message joins when one channel carries several at once.	Timestamps, speakers, name mentions, text and, when people use them, platform reply links carry the signal.	Reply-to link prediction plus clustering, structural graph models and LLM assignment do the work.	It is maturing: the best peer-reviewed model recovers whole conversations at a cluster F1 of 51.7 (Ma et al. 2022).
5. Segment topics inside a thread	It decides where a thread changes subject.	Word repetition, reply structure and discourse cues carry the signal.	Lexical cohesion (TextTiling, LCTSeg), reply-graph segmentation, BERT coherence scoring and LLM prompts find the boundaries.	It is maturing, and how to score it is contested (When F1 Fails , arXiv 2512.17083).
6. Link topics and stories across threads	It decides which threads discuss the same matter, and which matter is new.	Entities, time, topic words and embeddings carry the signal.	Topic tracking, first story detection, event threading and BERTopic clustering group the threads.	It is open: the notes found no benchmark for linking threads across workplace channels.



● Solved ● Maturing ● Open ➔ Input ⚙ Process ⬢ Decision 📄 Output

Figure 8. Threading as a workflow. Source: the stage table above; each stage's colour follows the maturity evidence cited in its row.

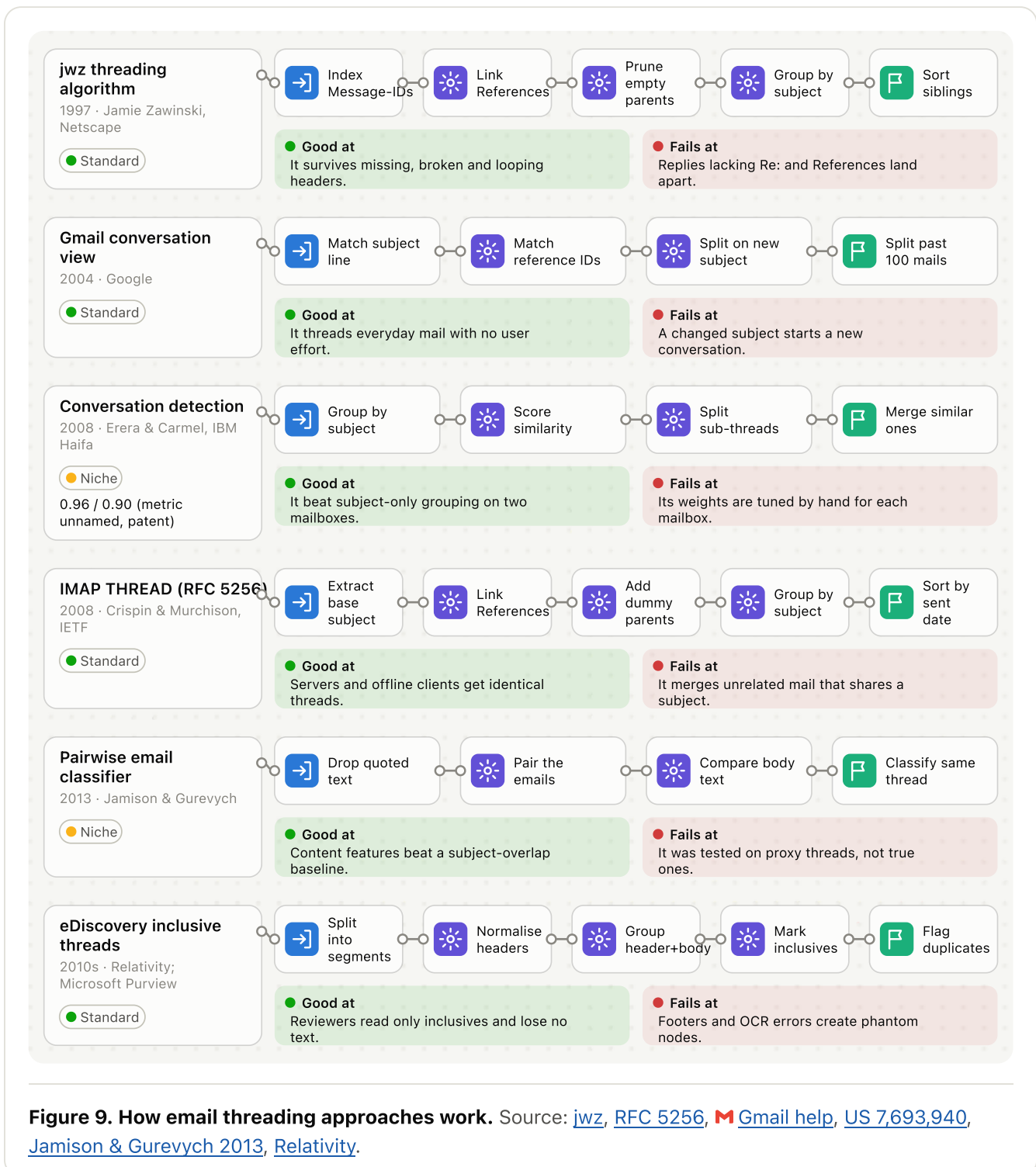
Email threading: the approaches

- Header threading was settled between [1982](#) and [2008](#). The content methods that try to work without headers ran from [1997](#) to [2013](#) and never got a shared benchmark.
- The notes found no new email-threading standard, benchmark or LLM-based method for 2014-2026; the absence rests on two searches and is not proof.

Approach	Year	How it works (one sentence)	Good at	Fails at	Status	Source
Reply headers (RFC 822, 2822, 5322)	1982 to 2008	A reply names its parent's Message-ID in In-Reply-To and the chain of ancestors in References.	It gives an exact parent link whenever every client keeps the headers.	RFC 822 left In-Reply-To as free text, and Lotus Notes, cc:Mail and some Exchange servers strip the headers.	It is the standard.	RFC 5322; Mailman wiki
jwz algorithm	1997	It links References into a tree, adds empty placeholder parents for missing messages, and groups orphans by base subject .	It cannot loop on malicious input and tolerates missing messages; jwz says it served about ten million users over six years (self-reported).	A reply with neither "Re:" nor References breaks the subject step, and disagreeing References are resolved by processing order.	It is the de facto standard.	jwz
IMAP THREAD (RFC 5256)	2008	It makes jwz 's method a server standard (REFERENCES) and adds a subject-only mode (ORDEREDSUBJECT).	Servers and offline clients compute identical threads.	It inherits jwz 's habit of merging unrelated mail that shares a subject.	It is the standard.	RFC 5256
 Microsoft Thread-Index	No dated source was found.	Each reply copies its parent's binary index and appends five bytes, so a message's place in the tree is a prefix of its index.	It stores each message's place in the conversation directly in a header.	Repke & Krestel note that methods built on it do not handle arbitrary mail from other clients.	It is standard only inside  Microsoft's ecosystem.	MS-OXOMSG; Repke & Krestel 2018
 Gmail conversation view	2004	It groups mail by subject and reference headers; automated mail must also arrive within one week of an earlier message.	It threads everyday mail automatically from subjects and headers.	A changed subject, or a conversation of more than 100 emails, starts a new conversation.	It is the product standard.	Gmail help

Approach	Year	How it works (one sentence)	Good at	Fails at	Status	Source
🍏 Apple Mail heuristics	Described in 2012	🍏 Apple replaced header-only threading with heuristics on subject and sender to "gauge user intention".	Replies with a consistent "Re:" subject stay together, according to 🍏 Apple's response.	When Amazon SES rewrote Message-IDs in 2019, 🍏 Apple Mail split the threads.	It is a product variant.	Phabricator T1735 ; Discourse
IR text matching (Lewis & Knowles)	1997	It matches a message's text, including quoted text, against earlier messages.	The abstract reports "a significant level of threading effectiveness" without headers.	Its dataset and numbers could not be found, so it cannot be compared with later work.	It was superseded by hybrid methods.	Record
Conversation detection (Erera & Carmel, IBM)	2008	It groups by core subject, then splits and merges groups by a weighted score of subject, date, participants and text.	It scored 0.96 and 0.90 on two private mailboxes against 0.87 and 0.70 for subject-only grouping; the patent read does not name the metric.	Its weights are set by hand for each mailbox.	It is niche and industrial.	US 7,693,940
Pairwise classifier (Jamison & Gurevych)	2013	It decides whether two emails share a thread from the text of their non-quoted parts alone.	Content features beat a subject-overlap (Dice) baseline.	Style and structure features added nothing, and its Enron Threads Corpus is a subject-and-quote proxy, not true threads.	It is niche.	ACL Anthology R13-1042
eDiscovery inclusive threading (Relativity,  Purview)	2010s;  Microsoft bought Equivio in January 2015	It splits each email into its quoted parts, groups by headers and bodies, and flags "inclusive" messages that contain all earlier text.	Reviewers read only the inclusive messages and lose no unique content, because forks and edited quotes yield several inclusions.	Footers, Bates stamps and OCR errors create extra inclusions and phantom nodes, and shared mailboxes over-merge.	It is the standard in legal review.	Relativity ;  Purview

- **Failures come from lost or rewritten headers more than from the algorithms.** Servers strip headers, sending services rewrite Message-IDs, and people edit subjects ([Mailman wiki](#); [Discourse](#)).
- **Clients disagree by design.** Palme noted in [1998](#) that header and subject methods yield different threads, so one conversation looks different in every client.



Chat disentanglement: the approaches

- A chat channel carries several conversations at once; in Elsner & Charniak's IRC sample, 2.75 conversations were active on average ([Elsner & Charniak 2008](#)).

- Since 2019 the standard formulation predicts, for each message, the earlier message it replies to, or a link to itself when it starts a conversation. Messages joined by links form one conversation ([Kummerfeld et al. 2019](#)).

Approach	Year	How it works (one sentence)	Good at	Fails at	Status	Source
Speaker clustering (Aoki; Camtepe; Acar)	2003-2005	It clusters speakers, not messages, by turn-taking and shared words.	Aoki's 2006 study first measured concurrent conversations: 1.76 floors on average among 8-10 speakers.	It assumes one speaker stays in one conversation, but the average IRC speaker joins about 3.3.	It was superseded.	Elsner & Charniak 2008
Single-pass clustering (Shen et al.)	2006	Each message joins the most similar recent thread by tf-idf cosine, or opens a new one below a threshold.	It defined the task and showed that a time window gives the largest gain (best F 0.612).	It was tested on 16 small class chats with labels supplied by the authors of the messages.	It was superseded.	Shen et al. 2006
Pairwise classifier + clustering (Elsner & Charniak)	2008	A classifier judges "same conversation?" for message pairs under 129 seconds apart, and a greedy pass clusters the votes.	It released the first public annotated corpus and metrics; it scored one-to-one 40.6 against 53.0 for humans.	Without name-mention features it falls to "essentially baseline".	It was the standard baseline for ten years.	Elsner & Charniak 2008; 2010
Context expansion (Wang & Oard)	2009	It adds words from nearby and same-author messages to each message before clustering.	It gained 24% in F over Elsner & Charniak on the same test set.	Its authors say the results are close to the limit of what one test collection can measure.	It was superseded.	Wang & Oard 2009
Ubuntu name heuristic (Lowe et al.)	2015	A reply is a message that starts with a recent username, and its question is that user's last message within three minutes.	It produced 930,000 dialogues for training dialogue models.	Only about 10% of its conversations are exactly right (precision 10.8).	It was disproven as disentanglement .	Kummerfeld et al. 2019

Approach	Year	How it works (one sentence)	Good at	Fails at	Status	Source
Reply-link feed-forward (Kummerfeld et al.)	2019	A feed-forward network scores each earlier message as the parent, and connected components become conversations.	It came with 77,563 annotated messages, the benchmark ever since, and reached a link F1 of 72.3.	It recovers whole conversations at a cluster F1 of only 36.2.	Its dataset is the standard; its model was superseded.	Kummerfeld et al. 2019
Pointer network (Yu & Joty)	2020	It points from each message to a parent among the previous 50, where 97% of parents lie, and runs online.	It works message by message as chat arrives and reached a cluster F1 of 44.5.	Errors on self-links, which start new conversations, drive its cluster score.	It is the standard online baseline.	Yu & Joty 2020
Structural model (Ma et al.)	2022	It adds who-is-speaking and who-refers-to-whom graphs to a BERT encoder before predicting reply links.	It holds the best peer-reviewed supervised cluster F1, 51.7.	About half of the conversations are still not recovered exactly.	It is the supervised state of the art.	Ma et al. 2022
Fine-tuned GPT-4o Mini (Pal et al.)	2026	It fine-tunes a small LLM on 153 training logs and asks for all reply links of a whole log at once.	It matches supervised models on links (link F1 74.4).	It reaches a cluster F1 of only 45.3, below structural models, and out-of-the-box GPT-4o scored 0.	It is evidence, not a product.	Pal et al., LREC 2026
Dialogue-level LLM assignment (Takada & Mori)	2026	One LLM call per message asks whether it joins one of the candidate conversations or starts a new one, showing the next 50 messages too.	✦ Gemini 2.5 Pro reached a cluster F1 of 61.3, above every supervised model.	It is a single run in one workshop paper, it reads future messages so it cannot run live, and GPT-4.1 with the same method scores 47.5.	It is claimed state of the art, not replicated.	Takada & Mori 2026

How the scores moved on the [Ubuntu IRC test set](#) (cluster F1 counts a conversation only if it is recovered exactly):

System	Year	Cluster F1	Caveat
Ubuntu name heuristic (Lowe et al.)	2015, scored in 2019	8.9	The heuristic was built for dialogue training, which needs only conversation prefixes.
Elsner & Charniak model	2008, scored in 2019	15.5	It was trained on a different, smaller channel.
Feed-forward (Kummerfeld et al.)	2019	36.2	A 10-seed ensemble reached 38.0 .
DialBERT	2020	44.1	The figure is taken from Ma et al.'s comparison table.
Pointer network (Yu & Joty)	2020	44.5	It is the only online system in this table.
Structural model (Ma et al.)	2022	51.7	It is the best peer-reviewed supervised result.
Discourse graphs (Li et al.)	2023	48.9	It leads on most other metrics but not on this one.
GPT-4o, out of the box (Pal et al.)	2026	0	The whole log went into one prompt .
GPT-4o Mini, fine-tuned (Pal et al.)	2026	45.3	Its link F1 of 74.4 matches supervised models.
GPT-4.1 with dialogue-level assignment (Takada & Mori)	2026	47.5	It is a single run and uses 50 future messages.
✦ Gemini 2.5 Pro with dialogue-level assignment (Takada & Mori)	2026	61.3	It is a single run, uses 50 future messages and has not been replicated.

- Numbers from before 2019 used other corpora and metric variants, so they cannot be put on this scale ([Elsner & Charniak 2010](#)).
- Platforms now record reply links when people use them: Slack from [January 2017](#) and Discord from [July 2021](#). Kummerfeld et al. note that Slack threads give only one layer of branching (✗ [arXiv 1810.11118](#)).
- The notes found no study of how many workplace messages are posted outside threads, so disentanglement stays relevant for channel-level messages.

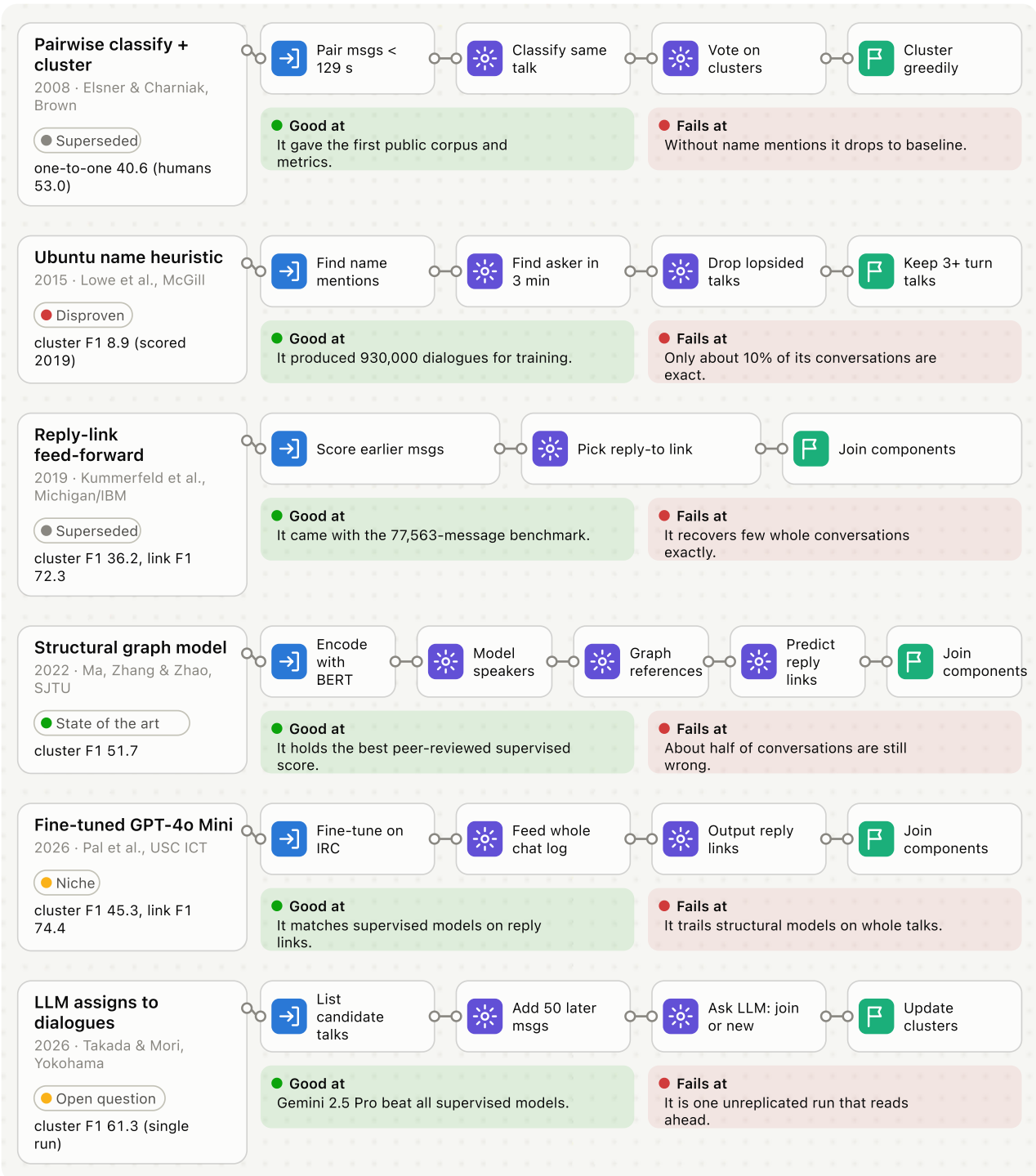


Figure 10. How chat disentanglement approaches work. Source: [Elsner & Charniak 2008](#), [Kummerfeld et al. 2019](#), [Ma et al. 2022](#), [Pal et al. 2026](#), [Takada & Mori 2026](#).

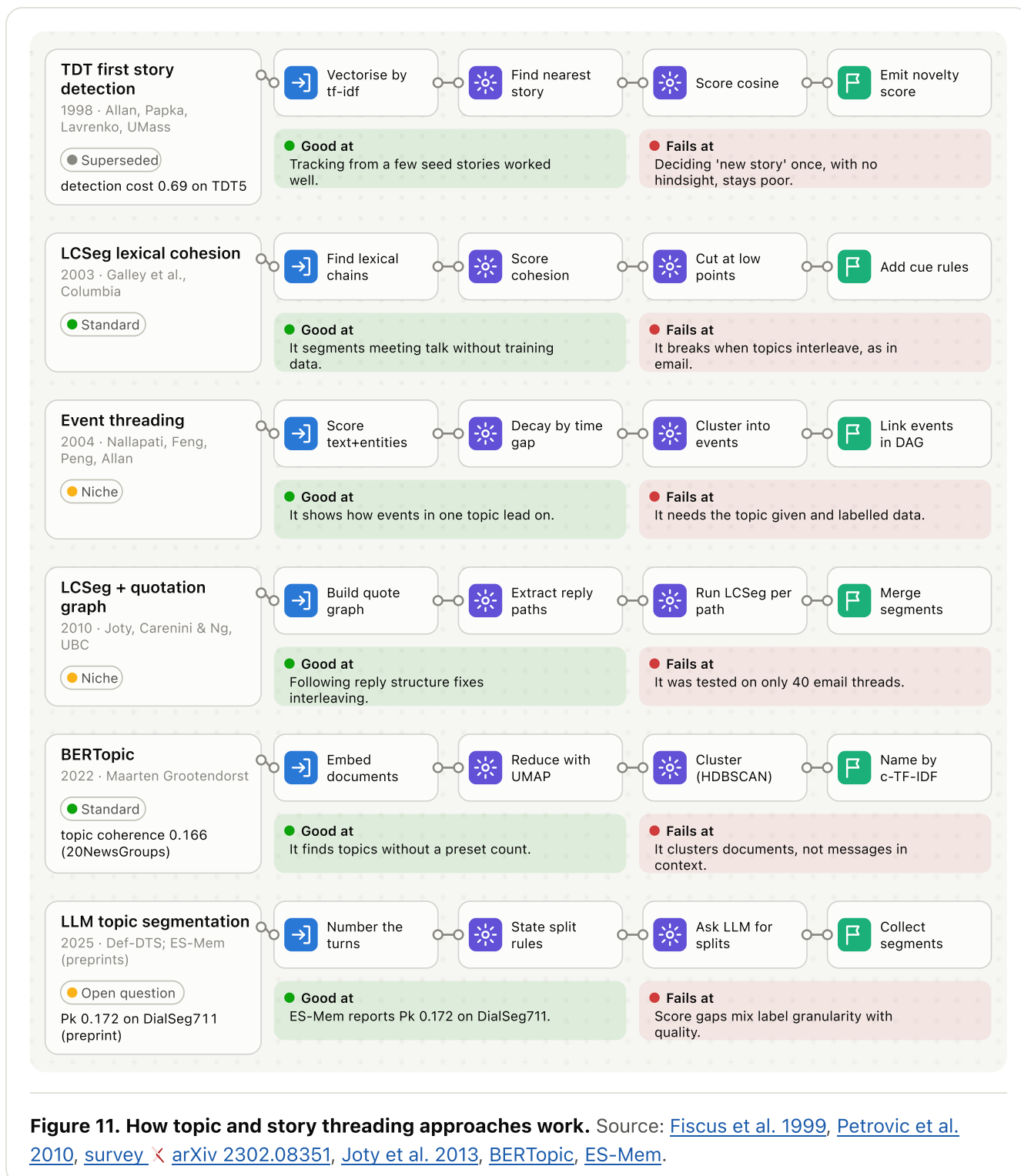
Above the thread: topics and stories

- Two problems sit above threading: splitting one thread where its topic changes, and linking threads that discuss the same matter.
- A 2026 study of an LLM assistant's chat log found that "a single thread can span multiple unrelated topics"; GPT-4.1 split its 815 threads into 1,267 segments, and nobody checked that split ([Nakamura & Eguchi, arXiv 2605.19627](#)). The log comes from people talking to a chatbot, not to each other.

Approach	Year	How it works (one sentence)	Good at	Fails at	Status	Source
TDT topic tracking and detection	1998	Systems follow a topic from a few example stories, or cluster a news stream into topics.	The best tracker missed 14% of stories at 0.2% false alarms, and four example stories instead of one improved tracking by 38% on average.	Automatic story boundaries raised detection cost by 97% on audio.	The programme was superseded; its methods live on.	Fiscus et al. 1999
First story detection (UMass)	1998	Each new story is compared with all earlier stories by tf-idf cosine, and its distance to the nearest one scores its novelty.	Petrovic et al. cut its run time on TDT5 from 28 hours to 2 with hashing, at similar cost.	Its detection cost stayed at 0.69 on TDT5, and on-line clustering scored 42% F1 against 82% for hindsight clustering.	The problem remains open.	Petrovic et al. 2010 ; Yang et al. 1998
Event threading (Nallapati et al.)	2004	It clusters stories inside one topic into events by text, entities and time, then links the events into a dependency graph.	It models how events inside one topic lead to each other.	It needs the topic in advance and labelled data; its scores were not retrieved.	It is niche but influential.	CIKM 2004 ; survey
Connect the dots (Shahaf & Guestrin)	2010	It finds the most coherent chain of articles between two given endpoint articles.	It produces storylines a reader can follow and check.	It needs both endpoints, its linear program was expensive, and it was evaluated with user studies; no benchmark numbers were found.	It is niche.	HUJI record ; survey
TextTiling (Hearst)	1994	It compares adjacent blocks of text and cuts where word overlap dips.	It needs no training and ships in NLTK.	It assumes topics follow one another in time order.	It is a standard baseline.	Hearst 1997

Approach	Year	How it works (one sentence)	Good at	Fails at	Status	Source
LCSeg (Galley et al.)	2003	It finds chains of repeated terms, scores cohesion between windows, and can add pause and cue-word rules for meetings.	Combining cues with cohesion gave "significant error reduction" on meeting transcripts.	It fails when topics interleave, as they do in email and blogs.	It is the standard unsupervised baseline for conversation.	ACL P03-1071 ; Joty et al. 2013
LCSeg on the quotation graph (Joty, Carenini & Ng)	2010	It runs LCSeg along each reply path of the fragment quotation graph and merges the segments.	It raised one-to-one agreement on email from 62.2 to 69.3, against 80.4 between humans (read from an extracted table, not re-checked).	It was tested on only 40 email and 20 blog conversations.	It is the reference for email; no later re-evaluation on email was found.	Joty et al. 2013
BERTopic (Grootendorst)	2022	It embeds documents, reduces dimensions, clusters them and names each cluster by its distinctive words.	It finds topics without fixing their number in advance.	It clusters documents, not messages in thread context, and its author calls its scores proxies for quality.	It is the standard practical tool.	arXiv 2203.05794
LLM segmentation (Def-DTS; ES-Mem)	2025	The LLM receives numbered turns and splitting rules and returns the segments.	ES-Mem reports Pk 0.172 on DialSeg711, against 0.500 for Def-DTS, with baselines copied rather than rerun (preprint).	A 2025 critique says such score gaps reflect annotation granularity as much as model quality.	The question is open.	arXiv 2601.07582 ; arXiv 2512.17083

- **No benchmark links threads across workplace channels.** The closest are within-channel disentanglement sets and TDT-style news linking across sources.



What was disproven or debated

- **Disproven: the Ubuntu Dialogue Corpus heuristic.** Kummerfeld et al. scored it against hand labels: its conversations reach a precision of 10.8 and a cluster F1 of 8.9, only 20% are right or a correct prefix, and 47% start on the wrong message (✗ [arXiv 1810.11118](#)).
 - Lowe et al. had built 930,000 dialogues with it in [2015](#) and left formal evaluation for later.
 - Dialogue training needs only conversation prefixes, so the corpus stays usable for that purpose.

- **Disproven: "one speaker, one conversation".** Early speaker-clustering systems assumed it, but Elsner & Charniak found the average speaker takes part in about 3.3 conversations ([Elsner & Charniak 2008](#)).
- **Debated: headers versus heuristics in email clients.** Netscape 4.0 dropped jwz's subject grouping and, by jwz's account, "never got threading working right" ([jwz](#)). 🍏 Apple went the other way, to subject-and-sender heuristics, and those split threads when a sending service rewrote Message-IDs in 2019 ([Phabricator](#); [Discourse](#)).
- **Debated: human agreement caps chat metrics.** Six annotators of the same IRC test set agreed with each other at only 52.98 one-to-one on average, ranging from 35.63 to 63.50, because some split conversations finely and others lump them ([Elsner & Charniak 2010](#)).
- **Debated: which chat metric to trust.**
 - Kummerfeld et al. replaced local agreement (loc3) and Shen-F with variation of information (VI), because loc3 makes strong assumptions about the data and Shen-F is asymmetric (✗ [arXiv 1810.11118](#)).
 - Cluster F1 is noisy: on the development set it varies about four times as much as link F1, so the DAG-LSTM authors chose checkpoints by link F1 ([Pappadopulo et al. 2021](#)).
 - Jiang et al.'s IRC F1 of 0.88 uses its own pairwise protocol and cannot be compared with exact-match scores around 0.4-0.5 ([Jiang et al. 2018](#); [Kummerfeld et al. 2019](#)).
- **Debated: Pk and WindowDiff for topic segmentation.**
 - Pevzner & Hearst showed in [2002](#) that Pk penalises missed boundaries more than extra ones and over-penalises near misses, and proposed WindowDiff.
 - Fournier showed in [2013](#) that Pk, WindowDiff and S all favour few or tightly clustered boundaries.
 - Joty et al. used one-to-one agreement instead, because Pk and WindowDiff assume linear text ([Joty et al. 2013](#)); a [December 2025](#) critique adds that annotation granularity drives many reported gaps.
- **Accepted as hard: first story detection.** Allan, Lavrenko and Jin titled their [CIKM 2000](#) paper "First Story Detection in TDT is Hard"; the notes could not retrieve its abstract.
 - The UMass detector's cost on TDT5 was 0.69 ([Petrovic et al. 2010](#)), and on-line event detection scored 42% F1 against 82% with hindsight ([Yang et al. 1998](#)).
 - Nothing disproved TDT itself: streaming detection and narrative extraction superseded it, and the survey that motivates narrative extraction criticises TDT for treating news as flat collections of events ([survey](#)).
- **Debated: whether LLMs can disentangle chat.** The results conflict because the experiments ask different questions.
 - Zero-shot GPT-3.5 picked the right parent message with a link F1 of 0.10, against 0.72 for a supervised model, on 100 sampled messages ([Li et al. 2023](#)).
 - Pal et al. gave GPT-4o the whole log in one prompt and got a cluster F1 of 0; fine-tuned GPT-4o Mini reached 45.3, below structural models ([LREC 2026](#)).
 - Takada & Mori asked, one call per message, which whole candidate conversation a message joins, and showed 50 later messages; ✦ Gemini 2.5 Pro reached 61.3 ([CEUR 2026](#)).
 - Prompt design explains most of the gap: on their development subset, the older pick-one-parent prompt scored 0.00 cluster F1 with GPT-4.1, and their prompt raised it to 31.43 ([CEUR 2026](#)). Model strength explains the rest, since GPT-4.1 with the new prompt reached 47.5 on the test set, below Ma et al.'s 51.7.
 - The winning result is a single run on 5,000 messages, uses future messages, and has no independent replication.

- **Debated: labels taken from threaded forums.** Kummerfeld et al. warn that stripping the structure from forums and mixing their conversations risks bias, because people write differently when structure is explicit ([✗ arXiv 1810.11118](#)).

Timeline

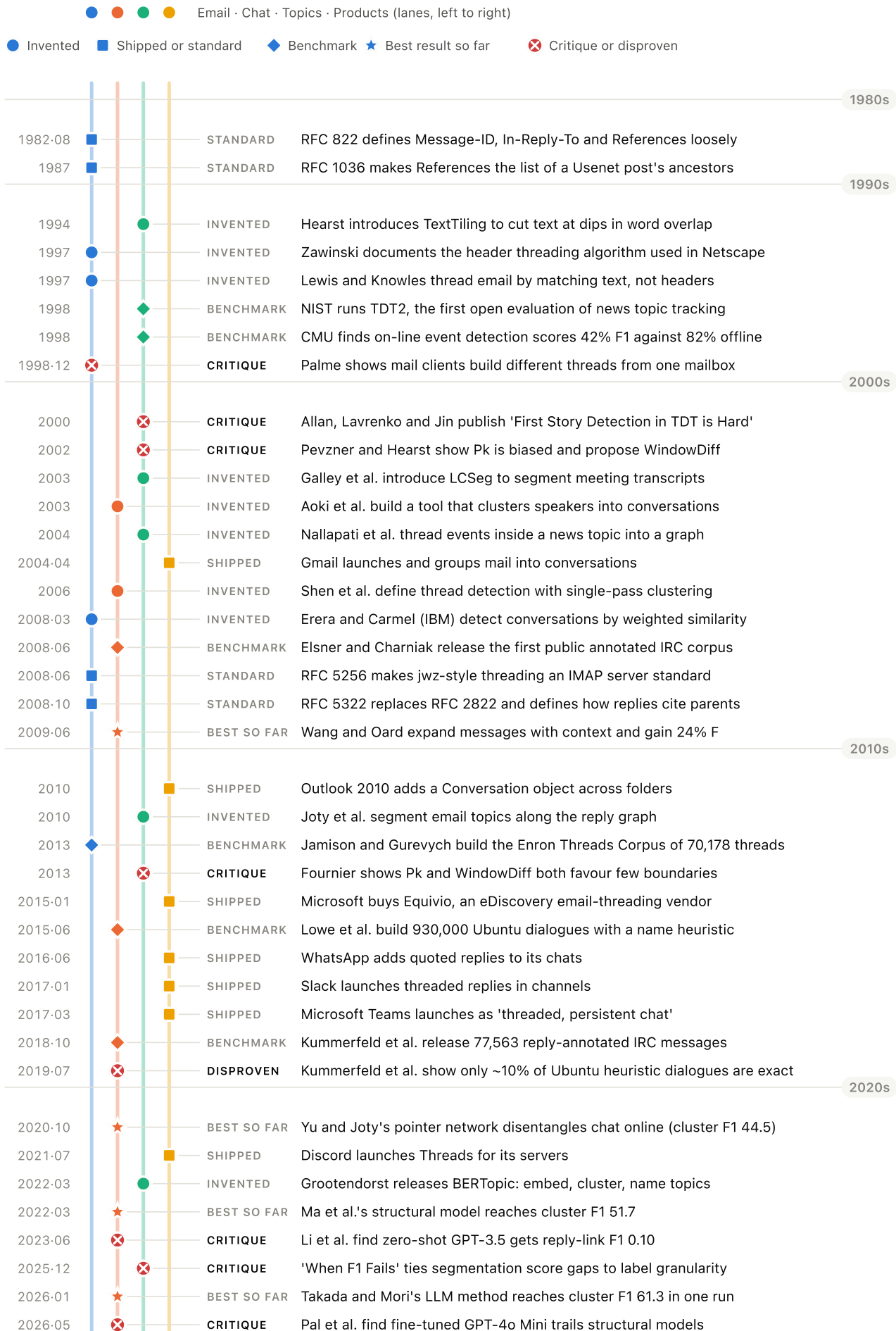


Figure 12. Timeline of threading research and products. Source: the three threading notes; every dated event carries its source in the figure's data file, `figures/threading_spec.json`.

- **1982-2008: email threading settled on headers.** RFC 822 named the reply headers in [1982](#), and RFC 5256 made jwz's algorithm a server standard in [2008](#).
- **1998-2004: news linking was measured.** TDT showed that tracking from examples works while first story detection stays hard ([Fiscus et al. 1999](#); [CIKM 2000](#)).
- **2008: chat disentanglement became measurable.** Elsner & Charniak released the first public annotated IRC corpus and its metrics ([ACL 2008](#)).
- **2017-2021: products shipped explicit reply threads.** Slack ([2017](#)) and Discord ([2021](#)) let people reply in threads, and 🟡 Microsoft Teams launched in [2017](#) as "threaded, persistent chat"; using threads stays optional.
- **2019 and 2026: the benchmark arrived, then the LLMs split it.** Kummerfeld et al.'s 77,563-message corpus refuted the Ubuntu heuristic in [2019](#); in [2026](#) one LLM result beat every supervised model while [another](#) fell short of them.

For a builder

- **Run threading as a pipeline with a confidence at each stage.** Keep every inferred link revisable, with the stage and signal that produced it, because later stages inherit its errors.
- **Trust headers and platform reply IDs first.** Use content similarity only as a fallback, and store the header link and the inferred link side by side.
- **For chat without reply links, predict reply links and then cluster.** If you use an LLM, ask which whole conversation a message joins, the only design that has beaten supervised models, in one unreplicated run. Measure link F1 and cluster F1 over several runs on your own labelled sample.
- **Segment topics along the reply structure, not along the timeline.** Joty et al. showed that time-ordered segmentation breaks when topics interleave ([JAIR 2013](#)).
- **Treat cross-thread links and "this is a new matter" decisions as suggestions for a person to confirm.** No benchmark measures them on workplace messages, and first story detection has stayed unreliable since TDT.

What shipped, what retreated, and how practitioners work

Industry history follows a clear rule: features that summarise or extract from templated mail became commodity, while features that need accountability (owner, due date, acceptance) retreated or remain unmeasured.

Templated extraction and constrained generation shipped.  Google's  Gmail extraction (Juicer) is the most rigorous production evidence, and it covers *machine-generated* mail.

Product	What it extracted	What happened	Source
 Gmail extraction ( Juicer)	Events and purchases from machine-generated mail; a rule-based extractor has run since 2013	Moving to ML raised event precision from 90.0 to 95.3 and purchase precision from 90.0 to 97.5 . Templates were whitelisted only when a weekly random sample scored above 90% precision; fixing faulty legacy labels added 8-9 precision points; more than 45,000 lines of rules were deleted	CIDR 2020
 Gmail extraction, privacy	Same	Serves over a billion users without engineers seeing individual emails	G Google Research
 Gmail schema.org markup	Structure supplied by senders themselves	Available to senders	G Google developers
 Smart Reply	Tightly constrained reply generation	Produced 10% of mobile replies in Inbox, showing that constrained generation scales	KDD 2016

- **None of this is evidence for extracting human commitments.**

Commitment and assistant features retreated.  Microsoft shipped commitment detection and then retreated.

Product	What it extracted	What happened	Source
 Cortana "Suggested Reminders" (2017)	Promises in sent mail	Shipped in 2017	eWeek
 Microsoft Briefing email	Outstanding commitments	Paused on 15 January 2023 , with no stated reason	 Microsoft support
 Cortana and Play My Emails	 Cortana's email feature	 Cortana reached end of support; Play My Emails was retired in June 2024	 Microsoft Support
 Google Inbox	Bundles, snooze, nudges,  Smart Reply	Closed in 2019 as part of product consolidation; snooze, nudges and  Smart Reply moved into Gmail, but bundles did not	w Wikipedia
x.ai	Assistant as a participant	Acquired by Bizzabo in June 2021	CB Insights
Clara Labs	Assistant as a participant	Acquisition dates conflict across sources	—
 Apple notification summaries	Summaries of notifications, including news	Falsely reported a news event in December 2024, so  Apple paused news summaries in iOS 18.3. Restored in the iOS 26 beta with the warning "Summarization may change the meaning of the original headlines". The fix was a UI disclaimer, not better accuracy	AppleInsider
 Gmail "AI Inbox" (January 2026)	Suggested to-dos	Starting with US English users (secondary sources only)	TechRepublic

- The assistant-as-participant model did not survive as a standalone business.

Meeting tools are commodity, and their accuracy is not measured.

Product	What it extracted	What happened	Source
 Zoom AI Companion (university review)	Action items	Accurate "when explicitly stated"	Seattle U
Otter (pilot)	Action items	"Sometimes incorrect"	Univ. of Alabama
LLM recap prototype (CSCW 2025 study, seven  Microsoft employees)	Meeting recaps	Users added, edited and deleted content; the study noted <u>hallucination risk on long dialogue</u>	 arXiv 2307.15793
 Microsoft Copilot, randomised trial (66 firms, 7,137 workers)	Vendor-reported impact, not accuracy	Regular users cut email time by about 3.6 hours a week; the <u>intent-to-treat effect was about 1.3 hours</u> ; only about 40% of those offered access used it regularly	 arXiv 2504.11436 , vendor-authored

- **No vendor publishes action-item precision or recall**, and the notes found no independent product study.

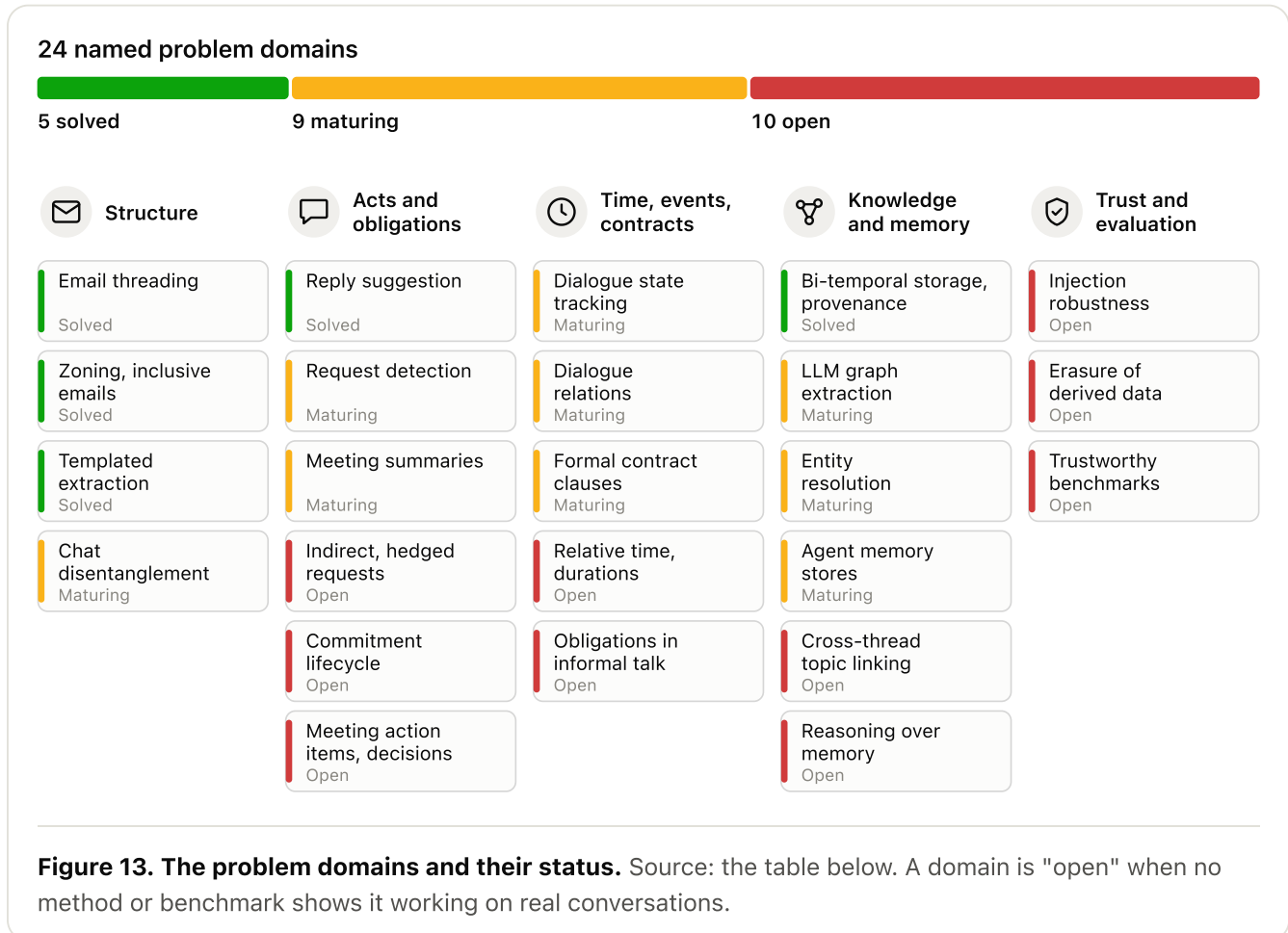
Two older industries already solved parts of the problem: eDiscovery, and enterprise graphs, which take two approaches.

Product	What it extracted	What happened	Source
eDiscovery ( Microsoft Purview, Relativity)	Treats a thread as a graph with "inclusive" terminal messages that contain all earlier text, so reviewers read only the minimal set	 Purview flags both branches of a fork as inclusive, where Outlook's conversation view could hide unique mid-thread content	 Purview docs; Relativity
 Palantir Ontology (enterprise graph)	Schema-first: humans define objects, links and actions	Data is mapped in rather than learned from text	 Palantir docs
 Atlassian Teamwork Graph (enterprise graph)	Built from tool metadata and permissions	Opened through an <u>MCP server</u> in May 2026	Nasdaq release
Contract-management vendors	Contract extraction	Publish no comparable extraction benchmarks; buyers are advised to run their own scored bake-off of about 200 contracts	Beri

- The notes found **no shipped product that tracks informal email and chat commitments across sources**, apart from  Microsoft's retreated features.

The solved, maturing and open map

The figure below names the problem domains of the field and marks each one; the table after it gives the evidence for each mark.

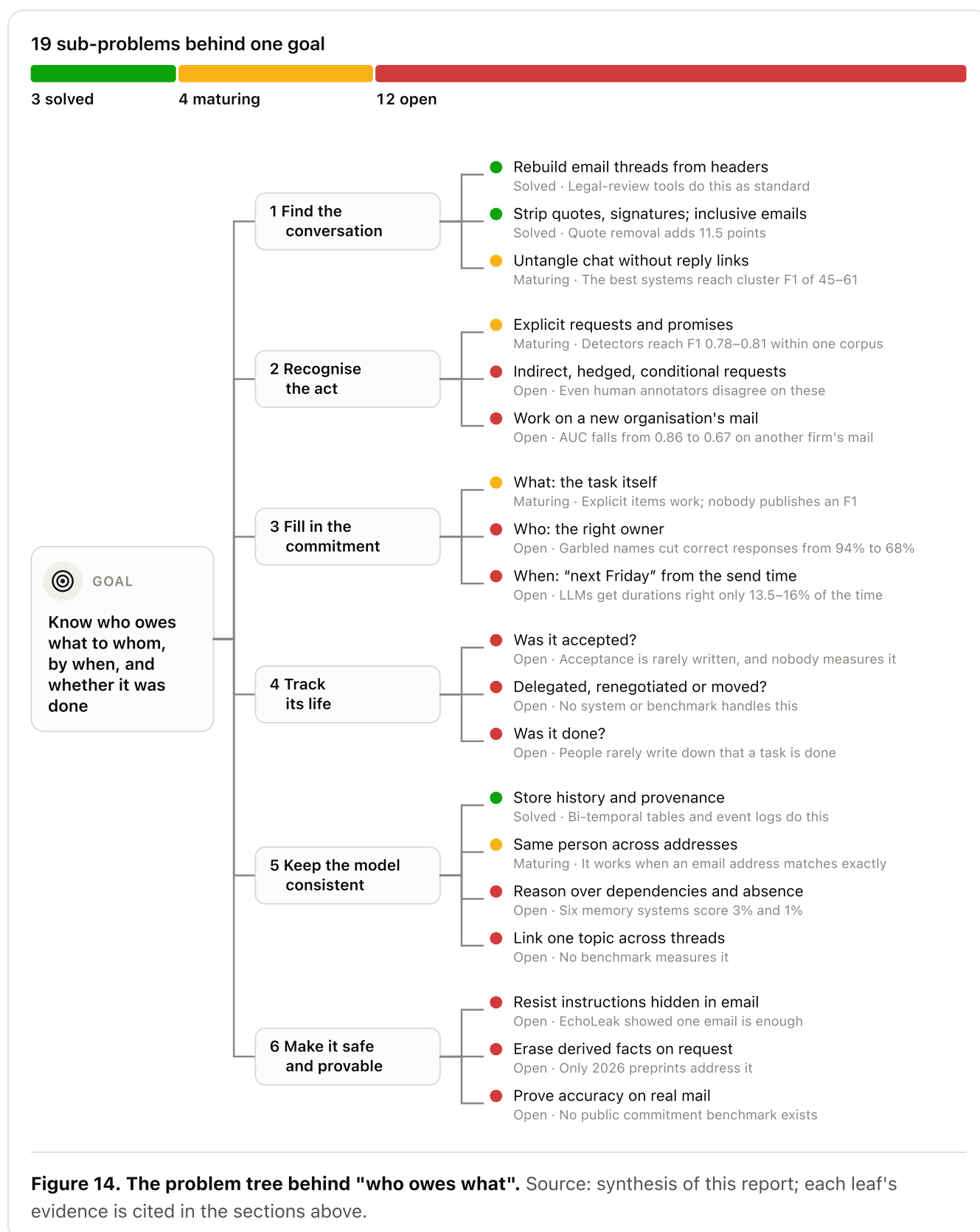


Status	Capability	Evidence and caveat
● Solved / industry standard	Header-based email threading, inclusive-email detection, near-duplicate grouping	Standard in Relativity and Purview (Purview); breaks when clients strip headers or servers insert footers
● Solved	Extraction from templated, machine-generated email (bills, bookings, purchases)	95-97% precision at Gmail scale (CIDR 2020); sender schema.org markup
● Solved	Transcription, meeting summaries, explicitly stated action items	Commodity in Zoom, Teams, Otter, Granola; accuracy unpublished, explicit items reported fine (Seattle U)
● Solved	Constrained reply suggestion	Smart Reply, 10% of Inbox mobile replies (KDD 2016)
● Solved (engineering)	Bi-temporal storage, event sourcing, provenance vocabularies, deterministic ER on strong keys, Fellegi-Sunter matching	SQL:2011 temporal tables; Fowler ; PROV-Q ; Splink

Status	Capability	Evidence and caveat
● Solved (research)	Email zoning as preprocessing	+11.5 accuracy points on requests (Lampert 2010); zoners brittle across corpora
● Maturing	Chat disentanglement without reply metadata	Cluster F1 roughly 45-61, single unreplicated LLM result at the top (CEUR)
● Maturing	LLM structured extraction into a graph; LLM entity matching	KGGen 66% fact capture, author-run (✗ arXiv 2502.09956); Peeters EDBT 2025
● Maturing	Agent memory stores (Mem0, Zep/Graphiti, Letta, MIRIX)	Benchmarks contested, see the "conflicting numbers" section below
● Maturing	Graph RAG for multi-hop and corpus summaries	Wins on reasoning and summary, ties on lookup, costs up to ~100× tokens (GraphRAG-Bench)
● Maturing	LLM thread and notification summaries	Shipped everywhere; 🍏 Apple's public failure; ROUGE does not track human judgment (EmailSum)
● Maturing	Contract clause extraction from formal documents	"Junior legal assistant" level (ContractEval)
● Maturing	Schema-bound dialogue state tracking	Close to solved on cleaned MultiWOZ; degrades on unseen entities (Qian 2021)
● Open	Commitment and request extraction from human email with lifecycle (accepted, delegated, discharged)	No LLM-era benchmark; cross-corpus AUC 0.86→0.67 (WSDM 2019); 🇺🇸 Microsoft Briefing paused
● Open	Meeting action items and decisions with correct owner and deadline	No verified LLM F1 on AMI or ICSI; labels derived, tiny (ICASSP 2023)
● Open	Indirect, conditional and hedged requests	Main driver of annotator disagreement (Lampert 2010)
● Open	Relative-time resolution against message timestamp; timelines and durations	No benchmark; durations 13.5-16% (Test of Time)
● Open	Cross-thread topic and story linking; new-topic detection	No benchmark; TDT found first-story detection unreliable (Allan, UMass)
● Open	Reasoning over dependencies, absence and deletion in memory	Cascade 3%, Absence 1% across six systems (MEME ; independence unchecked)
● Open	Erasure propagated through derived facts, summaries, embeddings	2026 preprints only; no regulator guidance (✗ arXiv 2602.17692)
● Open	Robustness to prompt injection from ingested messages	Production CVEs, architectural defences only (✗ arXiv 2509.10540)

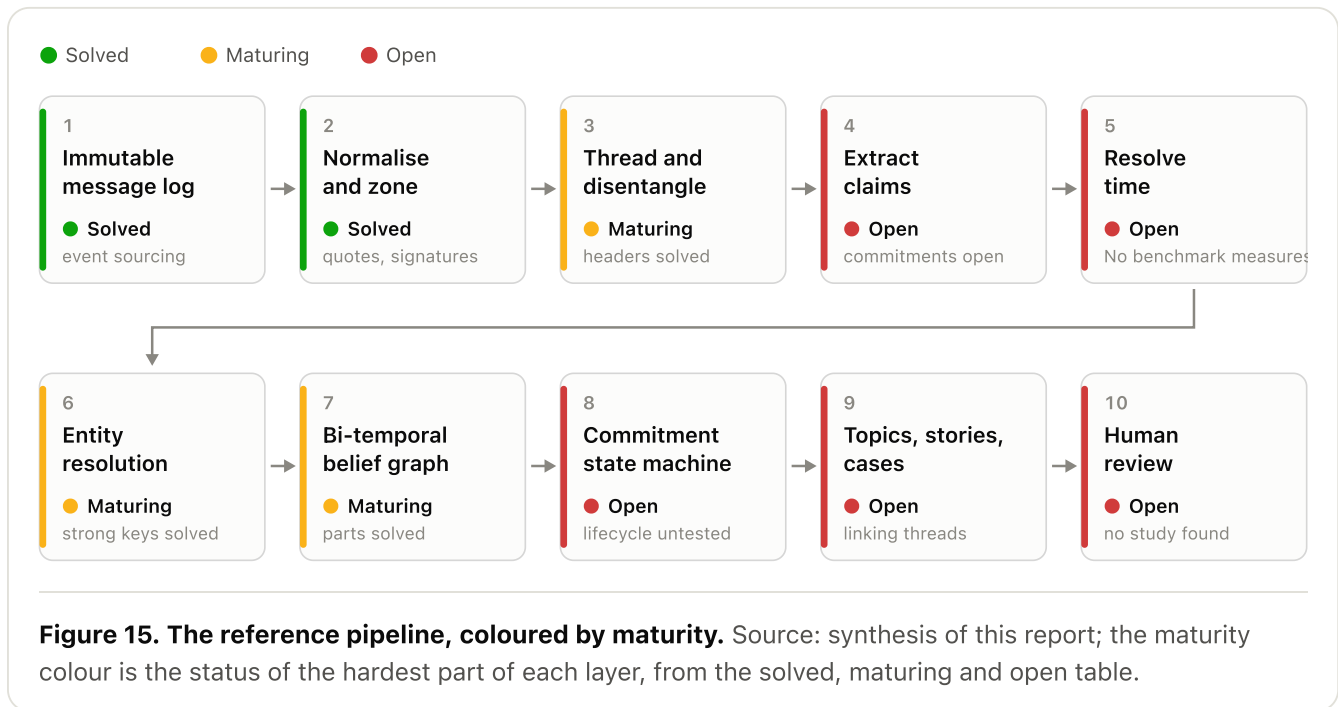
The hard problems have one thing in common. They need **pragmatic inference** (**intent**, **uptake**, **implicature**), **temporal reasoning anchored to context**, or **knowledge of what was *not* said**. Much of the world never reaches the channel: calls, hallway decisions and documents. "Nobody wrote 'done'" therefore tells you nothing about whether a commitment was kept.

The figure below breaks the goal most people actually want, a ledger of who owes what, into the sub-problems it needs. Most of the tree is open.



A reference architecture: log, claims, beliefs, review

This section is synthesis. Each layer is grounded in the sources cited, but the composition and several mechanisms (marked "inference") go beyond what any single source tested. The central design choice, inferred from the knowledge-representation notes, is to keep three layers apart. The first is the **immutable message**. The second is the **claim** extracted from it, a statement about a statement with provenance and confidence. The third is the **believed world state** computed from claims. RDF 1.2 "triple terms" with reifiers are the standards-track way to say "A claimed P" without asserting P. RDF 1.2 was still a Candidate Recommendation as of 7 April 2026 ([W3C RDF 1.2](#)). Property graphs get the same effect by promoting each claim to a node.



Layer	What it does	Grounding	Status
1. Immutable message log	Append-only store of raw messages, metadata, attachments and every extraction output (model id, prompt version , input hash)	Fowler: state is a replay of events; external query results must be logged so replays are deterministic (Fowler); PROV-O Entity/Activity/Agent (W3C)	Sourced pattern; treating LLM calls as "external queries" is inference
2. Normalise and zone	Strip quotes, signatures and footers; detect inclusive messages; deduplicate	Zoning +11.5 points (Lampert); inclusive sets (Relativity)	Sourced
3. Thread and disentangle	Native IDs first, header threading second, content disentanglement as a revisable cluster	Kummerfeld 2019 ; CEUR 2026	Sourced; "thread is not a primary key" is inference
4. Extract claims	LLM or classifier emits typed acts (request, commit, decision, fact, event) with span evidence, debtor/creditor , raw time expression and confidence	Purver role scheme (SRI); Singh commitments (AAAI-08); SEM Event/Actor/Place/Time (SEM)	Schema is inference from these

Layer	What it does	Grounding	Status
5. Resolve time deterministically	Anchor relative expressions to message send time and time zone; represent fuzzy times as intervals	OWL-Time intervals (W3C); LLM duration failures (ToT)	Inference (no benchmark exists)
6. Entity resolution	Auto-merge on strong keys (exact address), probabilistic plus review on weak keys (display names); merges stored as reversible sameAs assertions	Splink ; Senzing (vendor); Peeters	Reversibility is inference
7. Bi-temporal belief graph	Each fact has valid time and transaction time; supersession closes intervals instead of deleting; conflicts surface as flags; current belief is computed from justifications (TMS-style) so retracting a source retracts its consequences	Snodgrass bitemporal (w/ Wikipedia); Graphiti invalidation (GitHub); Doyle TMS vs AGM (Shapiro)	Sourced components; combination is inference
8. Commitment state machine	proposed → accepted/declined/unanswered → active → discharged (inferred vs confirmed) / cancelled / delegated; "violated" set by a clock-driven evaluator	Commitment machines (Singh group); Symboleo ; uptake (SEP)	States are inference
9. Topics, stories, cases	Thread (structural) → topic (semantic cluster) → story (TDT-style temporal cluster) → case or matter (human-curated, authoritative)	BERTopic ; TDT	Layering is inference
10. Human review	Confidence-gated queues; one-click confirm or dismiss; corrections logged as highest-authority events	Suchman contestability (EUSSET); users edit recaps (CSCW 2025)	Mostly inference; no HITL study on commitments found

Several engineering consequences follow from Fowler's event-sourcing rules. **Replays must never re-send real outgoing messages** (reminders, notifications), so the outbound gateway has to be disabled during rebuilds. Re-extraction with a better model should be an explicit new run rather than silent drift, with [idempotency](#) keyed on (message id, extractor version), which is an inference ([Fowler](#)). "We moved the meeting" is a **supersession, not a contradiction**: the new claim closes the old valid-time interval and both remain queryable "as known on date X". A real contradiction, where two sources disagree about the same time, should be flagged rather than silently overwritten. Store extraction confidence separately from source reliability rather than multiplying them into one number. The notes did not verify how well-calibrated LLM confidence is.

Evaluation has to be built in, because the literature cannot supply the numbers for your domain. The notes propose, as inference, measuring:

- [precision](#) per fact type, on samples shown to annotators with provenance
- [recall](#), from fully annotated threads
- commitment-state accuracy at a given time
- freshness lag
- count of unresolved conflicts
- [B-cubed](#) entity-resolution F1, plus the rate at which merges are reversed

Google's practice of whitelisting an extractor only when a weekly random sample clears 90% precision is a proven operational gate worth copying ([CIDR 2020](#)).

Ingested messages are an attack surface and a liability

Every inbound email is untrusted input to a privileged system. Greshake et al. named indirect prompt injection in 2023: instructions embedded in retrieved content can steal data, spread like a worm and contaminate ecosystems ([✗ arXiv 2302.12173](#)). EchoLeak (CVE-2025-32711, CVSS 9.3) made it zero-click against Microsoft 365 Copilot through one crafted email. It was fixed server-side in June 2025, and Microsoft reported no exploitation in the wild ([The Hacker News](#)). The exploit chain evaded the injection classifier with plain, human-sounding wording. It used reference-style Markdown to get past link redaction, auto-fetched images, and a Teams proxy the content security policy allowed. The proposed defences are architectural: prompt partitioning, provenance-based access control, strict CSP and least privilege ([arXiv 2509.10540](#)). ♦ Gemini's Gmail summaries could be made to show an attacker's fake security warning hidden in zero-size white text ([ODIN](#)). Slack AI had a prompt-injection path to private-channel data ([The Register](#)). A commitment tracker that reads inbound mail, holds privileged context and can remind or act has exactly the EchoLeak shape.

Memory poisoning makes injection persistent. SpAIware planted lasting instructions in ChatGPT's macOS memory for continuous exfiltration across future chats ([The Hacker News](#)). MINJA showed attackers can plant malicious memory records through ordinary queries, without write access, with over 95% injection success, and that those records affect other users' later queries ([✗ arXiv 2503.03704](#)). AgentPoison reached over 80% average attack success with under 0.1% of the store poisoned, though it needs write access ([arXiv 2407.12784](#)). For a world model, a forged email saying "the deadline moved to never" is a belief-revision attack. Mitigations implied by the evidence include:

- treat extracted items as untrusted drafts
- never auto-fetch or render links from ingested content
- isolate memory per tenant and per user
- weight claims by source trust
- require human confirmation before a claim from an external sender supersedes an internal one

The last two are inference.

Erasure of derived data is unsolved. No regulator has said how GDPR Article 17 applies to derived knowledge graphs, summaries or embeddings. EDPB Opinion 28/2024 says a model is anonymous only if identification is insignificantly likely by "all means" (secondary summaries only) ([CMS](#)). A 2026 preprint models agent memory as a dependency graph from raw memories through summaries to KG entities, so deletion can cascade. It documents "backflow", where deleted facts resurface through derived artifacts ([✗ arXiv 2602.17692](#)). The GONE benchmark measures multi-hop leakage after KG fact unlearning ([arXiv 2603.12275](#)). The architecture above makes deletion cascades possible only if every derived fact keeps a provenance edge back to its source messages. Bi-temporal "invalidate, never delete" conflicts with erasure, so a separate hard-delete path is needed. That last point is inference.

Workplace context adds legal and social limits. The EU AI Act bans inferring emotion from biometric data in workplaces; text-only sentiment is arguably outside the ban ([FPF](#)). In Germany, tools capable of monitoring behaviour trigger works-council co-determination ([Reusch Law](#)). Otter faces a consolidated class action, so far only

allegations, over recording without all-party consent ([NPR/Boise State](#)).  Microsoft's Productivity Score removed per-user names within weeks of launch after surveillance criticism ([PCWorld](#)). Suchman's 1993 critique describes the same backlash.

Things you probably have not thought of

A request is not a commitment, and silence is not acceptance. Speech-act uptake and Castelfranchi's acceptance condition mean every directive needs a state of proposed, accepted, declined or unanswered. Clark's grounding means "sounds good" may not cover everything in the thread. Discharge is rarely written down, so "done" should have inferred and confirmed variants, and the model needs an explicit "not observed" value for off-channel outcomes.

A who-knows-what index may be worth more than the fact graph. Transactive-memory research shows that groups retrieve by asking the right person ([Wegner 1991](#)). Authorship and reply patterns already reveal owners and experts per topic. Enron studies recover hierarchy from volume, response time and centrality, validated against a real org chart for one company only ([CorpRank](#)). Answering "ask whom" sidesteps many extraction-accuracy problems.

Inclusive emails eliminate most duplicate extraction. A reply quotes its whole history. Without eDiscovery-style inclusive detection, the same commitment is extracted once per reply and entity resolution has to clean it up later. Two related traps: mail servers insert footers that break body matching ([Relativity](#)), and date-filtered searches combined with threading can silently drop messages ([Above the Law](#)).

Benchmark numbers for the same product disagree by 20 points or more.

- In Mem0's own paper, full-context gpt-4o-mini (72.9) beat Mem0 (66.9) on LoCoMo, and Zep scored 66.0. Zep's re-run of Zep scored 75.1 and listed harness errors, including both speakers given the user role ([Zep blog](#)).
- Mem0's 2026 page reports 92.5 and concedes the answer model and judge can move scores by about 20 points ([Mem0](#)).
- A competitor's audit found **6.4% of LoCoMo's answer key wrong**, and a gpt-4o-mini judge accepting 62.8% of deliberately wrong but on-topic answers ([Penfield Labs](#)).
-  Redis, which sells its own memory product, reproduced  Vertex Memory Bank at 42.4 against roughly 70 reported ([Redis](#)).
- LoCoMo has no knowledge-update questions, so it does not test the supersession that bi-temporal graphs exist to handle ([Zep blog](#)).

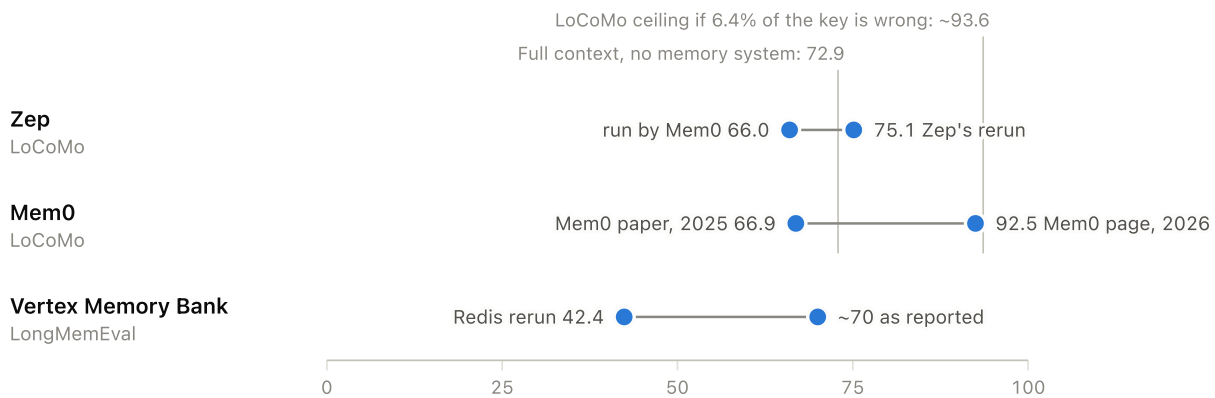


Figure 16. The same memory system, scored by different people. Sources: [Zep](#) and [Mem0](#) (both vendors), [Redis](#) (vendor rerunning others), [Penfield Labs audit](#) (competitor). Mem0's two scores are its own, a year apart.

Trust only same-harness reruns on your own data.

More memory can make answers worse. Mem0 reports that its multi-session score fell from 93.2 to 88.0 as more memories were retrieved (vendor-reported) ([Mem0](#)). Graphiti's invalidate-only design has no forgetting mechanism: schedule details and identity traits are structurally identical edges ([analysis](#)). Salience and decay need their own design.

Your training and evaluation data is probably contaminated or unrepresentative. Enron is one company in 2001, widely included in pretraining, and full of cleaning artefacts. Avocado cannot be redistributed. Commitment models lose much of their accuracy between the two ([Azarbonyad](#)). Budget for in-domain annotation, and expect a human-agreement ceiling around kappa 0.7. Fixing your own labels can be worth more than changing models, as [Google's 8-9 point gain from cleaning legacy labels](#) shows.

ASR errors become ownership errors. Corrupting names cut a meeting agent's response rate from 94.3% to 68% ([Meeting Delegate](#)). Without speaker diarisation and name verification, action items get assigned to the wrong person; one anecdote reports an item assigned to someone who did not attend ([Jamie blog](#)).

Users stop checking. A CHI 2025 study, seen only as a snippet, found participants often accepting AI outputs as accurate without verifying them ([ACM DL](#)). Review queues decay into rubber stamps unless confidence and evidence are shown at the moment of decision.

Valid tool calls are not completed tasks. EmailBench's 99.7% successful tool calls against a 33.5% scenario pass rate is a warning against operational metrics that look healthy while the world model is wrong ([arXiv 2609.31906](#), preprint).

The schema is political. Suchman's question of who defines the categories, and the Productivity Score backlash, both say: show people the commitments attributed to them, let them contest, and avoid per-person scoring by default.

Navigating the field

No single discipline owns this problem, so the vocabulary changes at every boundary. The literature is split across NLP and dialogue (ACL, EMNLP, NAACL, SIGDIAL, LREC), information retrieval and data management (SIGIR, WSDM, KDD, CIDR, VLDB, EDBT), semantic web (ESWC, Journal of Web Semantics, W3C), multi-agent systems (IJCAI, AAMAS-family commitment work), requirements engineering (RE, SoSyM for Symboleo), and HCI and CSCW (CHI, CSCW, ECSCW). Useful search terms include:

- **Speech acts and commitments:** "email speech acts", "commitment detection in email", "action item detection", "commitment machine", "social commitments multiagent"
- **Chat structure:** "conversation disentanglement", "email thread reconstruction", "email zoning"
- **Dialogue and events:** "dialogue relation extraction", "dialogue state tracking", "cross-document event coreference", "TIMEX3 normalisation", "topic detection and tracking / first story detection"
- **Storage and belief:** "bitemporal / valid time transaction time", "event sourcing projection", "temporal knowledge graph", "agent memory", "truth maintenance", "AGM belief revision", "RDF 1.2 triple terms"
- **Identity:** "entity resolution Fellegi-Sunter blocking"
- **Legal review:** "inclusive email eDiscovery"

Datasets	What it gives you	Caveat
Enron (Klimt & Yang)	~200K cleaned messages, 158 users	Single 2001 company; counts vary by version; pretraining contamination
Avocado (via Azarbonyad)	~940K emails, 279 accounts	LDC licence, not redistributable
BC3 (Ulrich 2008)	40 W3C threads with summaries and speech-act labels	Mailing list, not private workplace
EmailSum (ACL 2021)	2,549 threads with short and long summaries	ROUGE weakly tracks humans
Ubuntu IRC disentanglement (Kummerfeld 2019)	77,563 reply-linked chat messages	Technical domain
AMI / ICSI (Liu 2023)	Meetings; 381 derived action items on AMI	ICSI action labels not public
QMSum Mistake, MeetingBank (arXiv 2407.11919 ; Zenodo)	Summary-error labels; 1,366 council meetings	Summaries, not action items
DialogRE, MultiWOZ 2.4 (DialogRE ; MultiWOZ 2.4)	Dialogue relations; task-oriented state	Not workplace
CUAD, ContractNLI (CUAD ; ContractNLI)	Formal contract clauses and entailment	Formal text only
LoCoMo, LongMemEval (LongMemEval)	Long-memory QA	Contested scoring; few update questions

Surveys worth reading:

- Bian's LLM knowledge-graph construction survey (✗ [arXiv 2510.20345](#))
- An LLMs-and-knowledge-graphs survey (✗ [arXiv 2406.08223](#))
- Temporal KG completion surveys ([IJCAI 2023](#)), though these are mostly about predicting facts rather than maintaining extracted ones
- A 2026 belief-change survey the notes could not vouch for (✗ [arXiv 2608.14567](#))

The notes found no survey of conversational information extraction as a whole.

A suggested reading order for a builder moves from what to extract, to how to store it, to what has been measured:

1. **What to extract.** [SEP on speech acts](#), then [Clark & Brennan](#), then Suchman's critique ([EUSSET](#)), then [Singh on commitments](#).
2. **How it has been measured.** [Cohen 2004](#), [Lampert 2010](#), [Azarbondyad 2019](#) and [Kummerfeld 2019](#).
3. **How to store it.** [Fowler on event sourcing](#), bitemporal modelling, [PROV-O](#), [RDF 1.2](#), then [Zep/Graphiti](#), then the [Doyle/AGM overview](#).
4. **What production taught.**  Google's [CIDR 2020 Juicer paper](#), then the [GraphRAG-Bench](#) and [EchoLeak analysis](#).

Conclusion

The field's centre of gravity has moved from *classification* to *state maintenance*. The 2004-2019 work showed that [speech acts](#) and commitments can be detected moderately well inside one corpus. Industry has shown that accurate extraction is only possible where the text is templated. What nobody has demonstrated publicly is the next step: keeping a correct, current, contestable ledger of who owes what across email, chat and meetings, including acceptance, delegation and [discharge](#). That gap is an opportunity. It is also where  Microsoft retreated and where the benchmarks are weakest. A credible entrant therefore competes less on extraction accuracy than on **trustworthiness of state**, meaning [provenance](#) on every claim, [bi-temporal history](#), reversible merges, a clock that marks commitments violated, a deletion path that actually cascades, and a review loop people use.

Two strategic implications follow. Because no public benchmark measures commitment lifecycles, **a builder's own labelled evaluation set is a moat and a necessity**: per-tenant samples, Google-style weekly [precision gates](#), and task-level questions such as "what do I owe X?". And because the same ingestion pipeline that builds the world model is the injection vector, **security architecture cannot be bolted on later**. Untrusted senders' claims should never silently overwrite trusted state, and nothing extracted from an inbound message should trigger an outbound action without a [human in the loop](#).

What works and what does not

Some ideas have a track record, some have failed in public, and the most important ones are untested.

Ideas that work

Idea	Evidence	Source
Thread email by its reply headers (Message-ID, In-Reply-To, References).	Legal-review tools use it as standard.	 Purview
Strip quoted text and signatures before extracting anything.	Request accuracy rose from 72.3% to 83.8%.	Lampert 2010
Read only the "inclusive" email that contains a whole thread.	Legal-review tools use it as standard.	Relativity
Extract from templated, machine-generated mail with ML and a weekly precision gate.	 Gmail reaches 95–97% precision ( Google measuring its own system).	CIDR 2020
Fix the labels before changing the model.	 Google gained 8–9 points from cleaner labels; MultiWOZ trackers gained 7–10%.	CIDR 2020 , Qian 2021
Offer short, constrained suggestions instead of free generation.	 Smart Reply suggestions were used in 10% of Inbox mobile replies.	KDD 2016
Fill a fixed schema of slots from a conversation.	Trackers are close to solved on cleaned MultiWOZ data.	 arXiv 2104.00773
Use a graph for multi-hop questions and corpus summaries.	Graphs scored 53.4 vs 42.9 on reasoning and 64.4 vs 51.3 on summaries, at up to ~100× the tokens.	GraphRAG-Bench
Store facts bi-temporally on an event log.	It is standard engineering (SQL:2011, event sourcing).	Fowler
Match people probabilistically and auto-merge only on exact keys.	Splink is mature and scales past 100M records.	Splink
Transcribe and summarise meetings, keeping explicitly stated action items.	It is a commodity, and explicit items are reported as accurate.	Seattle U

Ideas that failed or retreated

Idea	What happened	Source
Make users label each message as a request or promise (The Coordinator).	It stayed in "fairly small and dedicated networks", and Suchman called it discipline and control.	 ACM Interactions , Suchman
Thread email by subject line alone.	Unrelated conversations get merged.	mail-threading docs
Untangle chat with simple heuristics (the Ubuntu Dialogue Corpus).	About 80% of its conversations were missing messages or had extra ones.	Kummerfeld 2019
Train a commitment detector at one company and use it at another.	AUC fell from 0.86 to 0.67.	WSDM 2019

Idea	What happened	Source
Remind people of commitments found in their email (📧 Cortana, the Briefing email).	📧 Microsoft paused it in January 2023 and gave no reason.	📧 Microsoft
Auto-sort the inbox into bundles (📧 Inbox by Gmail).	📧 Google closed Inbox in 2019, citing consolidation; bundles did not move into 📧 Gmail.	W Wikipedia
Publish AI summaries without checking them (🍏 Apple).	The summaries invented headlines; 🍏 Apple paused them, and the fix was a disclaimer.	AppleInsider
Score individual people's productivity (📊 Microsoft Productivity Score).	📊 Microsoft removed per-user names within weeks.	PCWorld
Let the LLM do the date arithmetic.	2024 models got durations right 13.5–16% of the time.	Test of Time
Put everything into a long context window.	11 of 13 models fell below half their short-context score at 32K tokens.	NoLiMa
Give an LLM the whole chat log to untangle at once.	Zero-shot GPT-4o scored cluster F1 0 ; assigning one message at a time scored 61.3 in a single, unreplicated run.	LREC 2026 , CEUR 2026
Trust a memory vendor's benchmark numbers.	Scores for one product vary by 20+ points, and 6.4% of LoCoMo's answer key is wrong.	Zep , Penfield
Treat incoming email as trusted input to an assistant.	EchoLeak: one crafted email made 📧 Microsoft 365 Copilot leak data (CVSS 9.3).	The Hacker News
Judge summaries by their ROUGE score.	ROUGE tracks human judgement only weakly.	EmailSum

Untested: no evidence either way

- Nobody has benchmarked how well LLMs extract commitments from real email.
- Nobody has measured whether a system can tell that a commitment was accepted, delegated or done.
- No benchmark links one topic across several threads.
- Only 2026 preprints ask whether deleting a message also erases what was derived from it.
- No study compares a [knowledge graph](#) with a million-token context on the same task.

Mind map of the report

The whole report on one page: eight branches, each leaf a finding with its status.



About this report and its session log

How it was made. One ✳ Claude Code desktop session on 25 October 2026, with Claude Opus 5.5 as coordinator, produced everything here. It followed the deep-research pattern:

- six researchers ran in parallel (✳ Claude Sonnet, medium effort), one per angle: theory and human practice, email and chat NLP, meetings and events and contracts, LLM memory systems, industry, and knowledge representation
- two more researchers filled the gaps (industry failures; LLM-era evidence and risks)
- one report writer synthesised the notes (✳ Claude Opus, high effort)

The coordinator then added the summary chapter, the figures and the infographics. The researchers' notes, with every source, are in [notes/](#).

Session log. The full transcript of that session, including every researcher's own transcript, is kept in the RESEARCH repository at:

LOG/0001-world-models-from-conversations/14ea11bf-dd20-4531-8b79-2fedea455aaa.jsonl

with the subagent transcripts in the folder of the same name beside it. The session id is 14ea11bf-dd20-4531-8b79-2fedea455aaa. LOG/ is not on  GitHub. A weekly job (tools/research_sessions.py in the repository) copies any newer version of the transcript into LOG/ and freezes LOG/ to cold storage under the label research-sessions.

To ask this session a question later, copy the .jsonl file into the  Claude Code project folder of the directory you start Claude in, then resume it by id. For the RESEARCH repository that folder is ~/.claude/projects/Users-annakubasiak-repos-RESEARCH/. Then run `claude --resume 14ea11bf-dd20-4531-8b79-2fedea455aaa` from the repository. The repository's LOG/README.md gives the same steps, and how to thaw the file from the freezer if the local copy is gone.

Glossary

Terms marked with a dotted underline in the text are defined here. On the web page, pointing at a term shows its definition.

AI systems

Agent memory

A store that lets an AI agent keep and retrieve information across conversations, such as facts about a user. Designs range from graphs to plain files. [Source](#)

BERT

A language model released by Google in 2018 that reads text in both directions. It is a standard base for classification tasks once fine-tuned. [Source](#)

Fine-tuning

Further training a pre-trained model on a smaller, task-specific dataset so it performs better on that task. [Source](#)

Human in the loop

A design in which a person checks or approves machine outputs, especially before consequential steps. Review queues work only if evidence and confidence are shown at the moment of decision. [Source](#)

LLM

Large language model: a neural network trained on huge amounts of text to predict and generate language, which can be prompted to summarise, classify or extract information. [Source](#)

MCP

Model Context Protocol: an open standard for connecting AI applications to external tools and data sources. An MCP server exposes a product's data to AI agents. [Source](#)

Prompt

The input text given to a language model, including instructions and any material to work on. Text from untrusted sources placed in a prompt can steer the model. [Source](#)

Token

The unit of text a language model reads and bills by, roughly three-quarters of a word. Token counts drive both cost and context limits. [Source](#)

Zero-shot

Asking a model to do a task with no worked examples and no task-specific training, relying only on instructions. [Source](#)

Architecture and data

AGM belief revision

A formal theory by Alchourron, Gardenfors and Makinson of how to change a set of beliefs on new information, keeping them consistent while giving up as little as possible. [Source](#)

Bi-temporal

Recording two time axes for every fact: when it was true in the world, and when the system learned it. This lets you ask what was believed on a given date. [Source](#)

Embedding

A list of numbers that represents a piece of text so that texts with similar meaning get similar numbers. Embeddings are used to search and group text, but they are hard to delete from. [Source](#)

Event sourcing

Storing every change as an immutable event in an append-only log, and computing the current state by replaying those events. Nothing is overwritten, so any past state can be rebuilt. [Source](#)

Idempotency

A property of an operation whose repeat has no further effect. Keying extraction on message id and extractor version means rerunning the same step does not create duplicates. [Source](#)

IMAP THREAD

An IMAP extension, RFC 5256 (2008), that lets a mail server return messages already grouped into threads, either by subject alone or by a version of the jwz algorithm. [Source](#)

Immutable log

A record to which entries are only added, never edited or removed. In this design it holds the raw messages and every extraction output, so everything derived can be traced and rebuilt. [Source](#)

OWL-Time

A W3C vocabulary for describing instants, intervals and durations in knowledge graphs, including fuzzy times represented as intervals. [Source](#)

owl:sameAs

A web-ontology link asserting that two identifiers refer to the same thing. Storing entity merges as sameAs assertions makes them reversible, because the assertion can be removed. [Source](#)

Projection

A view of current state computed by processing the event log. If the rules change, the projection is thrown away and rebuilt from the log. [Source](#)

Property graph

A graph database model in which nodes and edges carry key-value properties. A claim can be stored as its own node linked to the thing it describes. [Source](#)

PROV-O

The W3C vocabulary for describing provenance in RDF, built around entities (things), activities (what happened) and agents (who did it). [Source](#)

Provenance

The record of where a piece of data came from and how it was produced. Each extracted claim should point back to its source message and the extractor version. [Source](#)

RDF

Resource Description Framework: the W3C standard for knowledge as subject-predicate-object triples such as (Anna, owes, report). RDF 1.2 is a newer version that adds triple terms. [Source](#)

Reification

Turning a statement into an object you can attach information to, such as its source or confidence. In RDF 1.2 a reifier names a statement so "A claimed P" can be stored without asserting P. [Source](#)

Replay

Re-running the stored events through the processing logic to rebuild state. Replays must not re-send real outgoing messages such as reminders. [Source](#)

RFC 5322

The internet standard for the format of email messages (2008), successor to RFC 2822 (2001) and RFC 822 (1982). It says a reply should carry In-Reply-To and References headers that point to its parent. [Source](#)

Schema.org markup

Structured tags that senders embed in an email or web page so software can read, for example, a booking time directly. This avoids guessing from the text. [Source](#)

Simple Event Model

A schema for describing events through four parts: the event itself, its actors, its place and its time. [Source](#)

SQL:2011 temporal tables

Built-in support in the SQL:2011 database standard for tables that record how rows change over time, using system-time and application-time periods. [Source](#)

Supersession

When a new fact replaces an older one because the world changed, as when a meeting moves. The old fact's valid-time interval is closed rather than deleted, so both stay queryable. [Source](#)

Thread-Index

Microsoft's binary email header for conversations: the first message gets a timestamp and a unique ID, and each reply appends five bytes, so a message's place in the thread is a prefix of its index. [Source](#)

Transaction time

The period during which the database held a given record as its current belief. It lets you reconstruct what the system believed at any past moment, including mistakes. [Source](#)

Triple term

In RDF 1.2, a statement that can itself be referred to without being asserted as true. It allows saying that a source claimed something while leaving open whether it holds. [Source](#)

Truth maintenance system

A system that records why each belief is held, so when a supporting fact is retracted its dependent conclusions are withdrawn automatically. Doyle introduced the idea. [Source](#)

Datasets

AMI

The AMI Meeting Corpus: about 100 hours of recorded meetings with transcripts and annotations. [Source](#)

Avocado

A corpus of emails and attachments from 279 accounts of a defunct information technology company, licensed through the Linguistic Data Consortium and not freely redistributable. [Source](#)

BC3

The British Columbia Conversation Corpus: 40 email threads from W3C mailing lists with human summaries and speech-act labels. [Source](#)

ContractEval

A study that tests 19 language models on CUAD clause extraction. [Source](#)

ContractNLI

A dataset that tests whether a contract text supports, contradicts or does not mention a stated claim. [Source](#)

CraigslistBargain

A dataset of negotiation dialogues between buyers and sellers over online listings, in which the outcome is an agreed price. [Source](#)

CUAD

Contract Understanding Atticus Dataset: 510 commercial contracts labelled for 41 types of clause by legal experts. [Source](#)

DialogRE

A dataset for finding relations between people and things in dialogue, built from TV-show transcripts. [Source](#)

EmailBench

A 2026 preprint benchmark of 206 scenarios in which AI agents must complete email tasks. The best configuration passed 33.5%. [Source](#)

EmailSum

A dataset of 2,549 email threads with short and long human-written summaries, for evaluating thread summarisation. [Source](#)

Valid time

The period during which a fact is true in the real world, such as the date range of a meeting that was scheduled for 3 pm. It is one of two time axes in a bi-temporal store. [Source](#)

Enron corpus

A public collection of real emails from Enron employees, made public during the investigations that followed the company's 2001 collapse. Cleaned versions hold about 200,000 messages from 158 users; it is one company in one era. [Source](#)

Enron Threads Corpus

70,178 email threads built in 2013 from the Enron mail by grouping on subject and matching quoted text. It is a proxy, because Enron itself carries no true thread labels. [Source](#)

GraphRAG-Bench

A benchmark comparing graph-based and plain retrieval across fact lookup, complex reasoning and summarisation tasks. [Source](#)

ICSI

The ICSI Meeting Corpus: about 70 hours of recorded research-group meetings from the International Computer Science Institute, with transcripts and annotations. [Source](#)

LoCoMo

A benchmark of very long conversations with questions about them, used to test conversational memory. A competitor audit found 6.4% of its answer key wrong. [Source](#)

LongMemEval

A benchmark for chat assistants' long-term memory, with questions that need facts recalled from long histories of conversations. [Source](#)

MeetingBank

A dataset of 1,366 city council meetings with transcripts and summaries. [Source](#)

MINE

A benchmark created by the KGGen authors that measures how many facts from a text a knowledge-graph extractor captures. Its test texts are encyclopedic articles. [Source](#)

MultiWOZ

A large set of human-written, task-oriented dialogues across several domains such as hotels and restaurants. Version 2.4 cleans the state labels. [Source](#)

NoLiMa

A long-context benchmark in which the question and the answer-bearing text share few words, so models cannot find it by keyword matching. [Source](#)

QMSum Mistake

A set of 200 machine-written meeting summaries from QMSum, labelled with nine types of error, used to test how well a model can detect summary mistakes. [Source](#)

Test of Time

A synthetic benchmark of questions on timelines, durations and temporal reasoning, used to show that LLMs are weak at time arithmetic. [Source](#)

Evaluation and metrics

AUC

Area under the ROC curve: the probability that a classifier ranks a random positive example above a random negative one. 0.5 is guessing and 1.0 is perfect. [Source](#)

B-cubed

A score for clustering that checks, for each item, how much of its predicted group belongs to its true group and vice versa. It is used here to score whether people or organisations were merged correctly. [Source](#)

Cluster F1

An F1 score computed on whole groups of messages, so a conversation counts as correct only if the system recovers it exactly. It is stricter than scoring individual reply links. [Source](#)

Data contamination

When test material was already seen during a model's training, so scores overstate real ability. Enron email is widely included in pretraining, which makes it risky as an evaluation set. [Source](#)

Detection cost

The error score used in TDT: a weighted mix of the miss rate and the false-alarm rate. Cmin is the lowest cost over all decision thresholds; lower is better. [Source](#)

Domain adaptation

Techniques for making a model trained on one kind of data work on another, such as one company's email versus another's. Adversarial versions train the model to ignore features that reveal the source. [Source](#)

F1

A single score that combines precision and recall into one number, the harmonic mean, running from 0 to 100 or 0 to 1. It drops if either one is low. [Source](#)

Intent-to-treat effect

In a trial, the average effect on everyone offered a treatment, whether or not they used it. It is smaller than the effect on actual users because many offered people never take it up. [Source](#)

Ubuntu IRC

Chat logs from Ubuntu's technical-support IRC channels. Kummerfeld et al. annotated 77,563 messages with who replies to whom, to train and test disentanglement. [Source](#)

Inter-annotator agreement

How often independent human labellers give the same label to the same item. It caps what any model can reasonably score, since the labels themselves are noisy. [Source](#)

Joint goal accuracy

In dialogue state tracking, the share of turns where every slot of the tracked state is exactly right. One wrong value makes the whole turn count as wrong. [Source](#)

Kappa

A score for how much two annotators agree beyond what chance would produce, where 1 is perfect agreement and 0 is chance level. Values near 0.7 are usually read as substantial but imperfect agreement. [Source](#)

Krippendorff's alpha

A measure of agreement among any number of annotators that handles missing labels and different kinds of data. Like kappa, it corrects for chance agreement, and 1 means perfect agreement. [Source](#)

Local agreement

A disentanglement score that checks, for each message and the few messages before it (usually three), whether the system agrees with an annotator on same versus different conversation. It rewards local boundaries only. [Source](#)

One-to-one overlap

A disentanglement score that pairs each true conversation with at most one predicted conversation and counts the messages they share. Human annotators agree only about 53% on it, which caps what any system can show. [Source](#)

Pk

A topic-segmentation error rate: the chance that two sentences a fixed distance apart are wrongly put in the same or in different segments. Lower is better; it punishes missed boundaries more than extra ones. [Source](#)

Precision

Of the items a system flagged as correct, the share that really are. High precision means few false alarms. [Source](#)

Recall

Of the items that really are correct, the share a system found. High recall means few misses. [Source](#)

ROUGE

A family of scores that rate a machine summary by counting words and phrases it shares with a human reference summary. It correlates only weakly with human judgment of summary quality. [Source](#)

Shen-F

A disentanglement score that matches each true conversation with its best predicted conversation and averages their F1, weighted by size. It is asymmetric, and Kummerfeld et al. treat it as superseded by VI. [Source](#)

Language and theory

Commitment machine

A finite-state machine, a diagram of allowed states and transitions, that tracks a commitment from creation to completion. Yolum and Singh used them so agents may take any move that keeps the commitments intact. [Source](#)

Common ground

The knowledge and assumptions that the people in a conversation share and know they share. Grounding is how it is built up; one message alone does not establish it. [Source](#)

Debtor and creditor

In a social commitment, the debtor is the party who owes the action and the creditor is the party to whom it is owed. A commitment record needs both, plus a context. [Source](#)

Dialogue act

A label for what an utterance does in a conversation, such as question, answer, suggestion or agreement. It is the dialogue-analysis counterpart of a speech act. [Source](#)

Directive

A speech act that tries to get the listener to do something, such as a request, order or instruction. Each directive needs a tracked state: proposed, accepted, declined or unanswered. [Source](#)

Discharge

To fulfil a commitment so that it no longer binds. People rarely write that a promise was kept, so a tracker must separate inferred discharge from confirmed discharge. [Source](#)

Discourse structure

How the parts of a conversation or text connect: which message answers which, which topic is still open and how one statement supports or contrasts with another. A new message usually attaches to a recent open topic. [Source](#)

Event-indexing model

Zwaan and Radvansky's account that people store events in memory indexed by time, space, entity, causation and intentionality. It suggests the dimensions a machine world model should track. [Source](#)

Variation of information

An information-theory measure of how far two groupings of the same messages differ. Disentanglement papers report it rescaled so that higher means closer agreement. [Source](#)

WindowDiff

A topic-segmentation error rate proposed in 2002 to fix Pk's biases by comparing boundary counts inside a sliding window. A 2013 analysis found that it is biased too. [Source](#)

Felicity conditions

The circumstances that must hold for a speech act to work, for example that the speaker has standing to order and the listener takes the order up. If they fail, the act misfires. [Source](#)

Focus stack

In Grosz and Sidner's theory, a stack of the topics currently in play, with the most recent on top. It explains why a new message usually attaches to a recently opened topic. [Source](#)

Grice's maxims

Four conversational rules that speakers normally follow: say enough but not too much, be truthful, be relevant and be clear. Listeners infer extra meaning when a speaker seems to break one. [Source](#)

Grounding

The process by which people building a conversation show each other that they understood what was said, to a level good enough for the current purpose. Email makes this slow, because acknowledgements often never arrive. [Source](#)

Illocutionary act

The act performed in saying something, such as warning, requesting or promising, as distinct from the words themselves. It is the layer of meaning that commitment extraction tries to recover. [Source](#)

Implicature

What a speaker means beyond the literal words, which the listener works out by assuming the speaker is being cooperative. "Some of the tasks are done" implicates that not all are. [Source](#)

Indirect speech act

A request or other act performed by saying something that looks like a different act, such as "Could you maybe look at this?" meaning "Please look at this". Pattern matching on the surface form tends to miss them. [Source](#)

Pragmatic inference

Working out what a speaker intends from context, not from the words alone, including intent, whether a request was accepted, and what was left unsaid. It is the common difficulty behind the open problems in this report. [Source](#)

Rhetorical Structure Theory

A theory that describes a text as a tree of parts joined by relations such as explanation, contrast or elaboration. It gives a way to say how one message relates to another. [Source](#)

SDRT

Segmented Discourse Representation Theory: a theory that links the pieces of a dialogue with typed rhetorical relations such as question-answer or correction. Unlike a simple tree, it can model several interleaved topics. [Source](#)

Sensemaking

The ongoing human process of turning scattered information into an understanding, by collecting evidence and revising a working picture. Pirolli and Card's version alternates between foraging for data and building a schema. [Source](#)

Situation model

The mental picture a reader builds of the events described, as opposed to a memory of the words. It is organised by time, place, people or objects, cause and intention. [Source](#)

Social commitment

In Munindar Singh's model, an obligation one party (the debtor) owes to another (the creditor) within a context. It is defined by what is publicly owed, not by anyone's private beliefs or intentions. [Source](#)

Law and policy

Biometric emotion recognition

Inferring a person's emotional state from body signals such as voice, face or gait. The AI Act bans it in workplaces and schools; the report notes text-only sentiment is arguably outside the ban. [Source](#)

Co-determination

A legal arrangement, strong in Germany, under which employees' representatives share decisions with management. Tools able to monitor behaviour trigger it. [Source](#)

Consideration

In contract law, something of value each side gives up or does, such as payment or a promise. With offer and acceptance, it helps decide whether a promise is legally binding. [Source](#)

eDiscovery

Electronic discovery: finding, collecting and reviewing electronic documents such as email as evidence in litigation or investigations. It has mature tooling for email threads and duplicates. [Source](#)

Speech act

Something a person does by saying it, such as promising, requesting, ordering or asserting, as opposed to merely conveying information. Speech acts are the reason a message counts as an action. [Source](#)

Speech act theory

The philosophical account, started by J. L. Austin, that treats utterances as actions with effects in the world. It lets a system ask what a message does and not only what it says. [Source](#)

Textbase

A mental record of the words and sentences of a text, as distinct from the situation they describe. Cognitive science contrasts it with the richer situation model. [Source](#)

Transactive memory

A group's shared memory system, in which members remember who knows what and ask the right person instead of storing everything themselves. It supports building a who-knows-what index. [Source](#)

Uptake

The listener's recognition and acceptance of what a speaker was trying to do. A request only becomes a commitment once there is evidence of uptake. [Source](#)

Utterance

A single stretch of speech or writing by one person, such as a sentence in a meeting or one chat message. It is the basic unit that speech-act and dialogue analysis labels. [Source](#)

EDPB

The European Data Protection Board, which brings EU national data-protection authorities together and issues guidance so the GDPR is applied consistently. [Source](#)

EDPB Opinion 28/2024

The Board's 2024 opinion on data protection in AI models. It says a model can count as anonymous only if extracting personal data from it is insignificantly likely by any reasonable means. [Source](#)

EU AI Act

The EU's Regulation 2024/1689, which sorts AI systems by risk and bans some uses outright, including inferring emotions from biometric data in workplaces. [Source](#)

GDPR

The EU General Data Protection Regulation (2016/679), which sets rules for processing personal data of people in the EU, including rights to access and erasure. [Source](#)

Novation

Replacing a party to a contract with a new party, or an old obligation with a new one, with everyone's agreement. The old party is released. [Source](#)

Right to erasure (Article 17)

Article 17 of the GDPR: a person's right to have their personal data deleted without undue delay under certain conditions. How it reaches summaries and graphs derived from messages is unsettled. [Source](#)

NLP tasks and methods

ASR

Automatic speech recognition: software that transcribes spoken audio into text. Its errors, especially on names, carry into everything built on the transcript. [Source](#)

Base subject

A subject line with reply and forward markers such as "Re:" and "Fwd:" and list tags removed. RFC 5256 defines exactly how to compute it, so that servers group threads alike. [Source](#)

BERTopic

A method that groups documents into topics by clustering their embeddings and then naming each cluster with its characteristic words. [Source](#)

Blocking

A speed-up in entity resolution that only compares records sharing a cheap key, such as the same surname, instead of every pair. It makes matching tractable on very large datasets. [Source](#)

Coreference

Recognising that different expressions refer to the same thing, for example "Anna", "she" and "the CEO". Cross-document event coreference does so for events across separate documents. [Source](#)

Deduplication

Removing repeated copies of the same content. In email, a reply quotes its whole history, so without it the same commitment would be extracted once per reply. [Source](#)

Dialogue state tracking

Keeping a running record of what a user wants in a task-oriented conversation, as values for fixed slots such as restaurant area or booking time. It updates as each turn arrives. [Source](#)

Diarisation

Working out who spoke when in a recording, so that each stretch of transcript is attributed to the right person. Without it, action items can be assigned to the wrong owner. [Source](#)

Symbolleo

A specification language for legal contracts that models obligations, powers and a contract's lifecycle as state machines, so compliance can be checked by software. [Source](#)

Works council

An elected body of employees at a workplace, common in Germany, with legal rights to be consulted or to agree before employers introduce tools that could monitor staff. [Source](#)

Disentanglement

Separating the interleaved conversations in one chat channel into individual threads. It is framed as predicting which earlier message each message replies to, then grouping. [Source](#)

Email zoning

Splitting an email body into zones such as new text, quoted earlier messages, forwarded content and signature. Classifiers then read only the new text, which raised request-detection accuracy in the cited study. [Source](#)

Entity resolution

Deciding which records or mentions refer to the same real-world person or organisation, and merging them. Errors either split one person into two or fuse two people into one. [Source](#)

Event threading

Grouping the news stories inside one topic into events and linking the events by which led to which. It produces a small graph of the story's development. [Source](#)

Fellegi-Sunter

A classic probabilistic method for record matching that weighs how strongly each agreeing or disagreeing field, such as name or address, indicates a true match. It can be trained without labelled examples. [Source](#)

First story detection

Deciding, as each story or message arrives, whether it reports a new event or continues a known one. The system decides once, without later context, and accuracy has stayed low. [Source](#)

Fragment quotation graph

A graph of the text fragments in an email thread that links each fragment to the fragments it quotes. It captures replies to parts of a message, which a message-level reply tree misses. [Source](#)

GraphRAG

A form of RAG that first builds a knowledge graph from the documents and retrieves over that graph and its summaries. It handles broad, multi-hop questions better but costs far more tokens. [Source](#)

Hallucination

When a language model states something that is false or not supported by its input, as if it were true. In summaries this means invented or distorted content. [Source](#)

Inclusive email

A message in a thread that already contains the text of all earlier messages, so reviewers can read it alone and skip the rest. Legal teams use this to cut review volume. [Source](#)

Information extraction

Automatically pulling structured facts, such as people, dates, events and relations, out of unstructured text. Applied to dialogue, it is called dialogue IE. [Source](#)

jwz algorithm

Jamie Zawinski's email threading method from Netscape Mail: it links messages through their References headers, adds placeholder parents for missing messages and groups leftovers by subject. RFC 5256 made it an IMAP standard. [Source](#)

Knowledge graph

A network of entities (people, places, things) joined by labelled relationships, stored so a program can query connections. Facts are nodes and edges rather than rows in a table. [Source](#)

LCSeg

A 2003 topic segmenter for meetings that tracks chains of repeated terms and cuts where they thin out, optionally adding pause and cue-word rules. It assumes topics follow one another in time order. [Source](#)

LLM-as-judge

Using one language model to grade another's answers instead of human reviewers. It is cheap, but the judge can be lenient, as the audit of LoCoMo showed. [Source](#)

Long context

Giving a language model a very large amount of text directly in its prompt instead of retrieving selected pieces. Accuracy tends to degrade as the context grows, especially for information in the middle. [Source](#)

Message-ID, In-Reply-To and References headers

Standard email header fields: Message-ID names a message, In-Reply-To names the one it answers, and References lists the earlier chain. Mail software uses them to rebuild threads exactly. [Source](#)

Multi-hop

A question that can only be answered by chaining several facts together, for example finding a person's manager's team. Graph methods help most on these. [Source](#)

Near-duplicate

A document whose text is almost identical to another, such as the same email with a different footer. Review tools group them so the same content is not handled twice. [Source](#)

OpenIE

Open information extraction: pulling relations out of text without a predefined schema of relation types. It is flexible, but the output is inconsistent from one document to another. [Source](#)

RAG

Retrieval-augmented generation: the system first looks up relevant passages from a store, then gives them to a language model to write the answer. This grounds answers in stored text. [Source](#)

Relation extraction

Finding pairs of entities in text and the relationship between them, such as who works for whom. Dialogue relation extraction does this inside conversations. [Source](#)

Reply-to link

A pointer from a chat message to the earlier message it answers. Models predict one per message; linked messages form one conversation, and a link to itself starts a new one. [Source](#)

Schema-based and schema-free extraction

Two ways to extract a knowledge graph. Schema-based extraction fills a predefined set of entity and relation types, which is precise but scales poorly; schema-free lets the model invent types, which is flexible but inconsistent. [Source](#)

Single-pass clustering

Clustering items in arrival order: each new item joins the most similar existing cluster if it is similar enough, or starts a new one. Earlier decisions are never revisited. [Source](#)

TextTiling

Hearst's 1994 topic-segmentation method: it compares the words of adjacent blocks of text and places a boundary where their overlap dips sharply. It needs no training data. [Source](#)

Threading

Grouping messages into conversations. For email it is done from header fields that point to the parent message; without them, systems fall back on subject lines or content similarity. [Source](#)

TIMEX3 normalisation

Converting time phrases such as "next Friday" into standard calendar values, using the TIMEX3 annotation scheme of TimeML. Relative phrases need the message timestamp to resolve. [Source](#)

Topic detection and tracking

A research programme on finding new topics in a news stream, following them over time and spotting the first story about each. It found first-story detection unreliable. [Source](#)

Topic segmentation

Splitting a text, meeting or thread into stretches that each discuss one topic. Most methods look for places where the vocabulary changes. [Source](#)

Topic tracking

Following one topic through a stream of stories after being shown a few examples of it. In the TDT evaluations it was the task that worked best. [Source](#)

Security

Attack surface

All the points where an attacker could feed input to or extract data from a system. Every inbound email is one for a mail-reading assistant. [Source](#)

Backflow

The report's term (from a 2026 preprint) for deleted facts reappearing through derived artifacts such as summaries or graph entities that were built from them. [Source](#)

Content security policy

A browser mechanism that restricts which sources a web page may load from, intended to limit data leaks and injected code. [Source](#)

CVE

Common Vulnerabilities and Exposures: a public catalogue that gives each known security flaw a unique identifier, such as CVE-2025-32711, so everyone refers to the same issue. [Source](#)

CVSS

Common Vulnerability Scoring System: a standard 0-10 severity rating for security flaws. 9.0 and above is Critical. [Source](#)

Data exfiltration

Unauthorised transfer of data out of a system to an attacker. In prompt-injection attacks, the model is tricked into leaking private data, for instance through a link or image request. [Source](#)

Indirect prompt injection

Prompt injection where the malicious instructions are hidden in content the model retrieves, such as an incoming email or web page, rather than typed by the user. Greshake et al. named it in 2023. [Source](#)

Least privilege

The principle that a user or program should hold only the access needed for its task, so a compromise can do less damage. [Source](#)

Machine unlearning

Making a trained model or knowledge store forget specified data, as if it had never been learned. The GONE benchmark tests whether related facts still leak. [Source](#)

Memory poisoning

Planting false or malicious entries in an AI system's long-term memory so they persist and shape later answers, including those for other users. [Source](#)

Prompt injection

An attack in which text fed to a language model carries instructions that override or redirect what the model was told to do. [Source](#)

Zero-click

An attack that works without the victim doing anything, such as opening or clicking. EchoLeak needed only that a crafted email arrive and the assistant process it. [Source](#)