

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. **Header-based threading, quoted-text deduplication, templated extraction from machine-generated email** ([G](#) Google reports 95-97% precision) and **meeting transcripts with summaries are commodity. Human-to-human commitments are not.** 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, and 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 together give a precise vocabulary for what to extract and how to store it. The engineering lesson that runs through the evidence is that 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, carry its confidence and extractor version, and be invalidated rather than deleted. A human should be able to contest it. 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. And no regulator has yet said how GDPR erasure reaches derived graphs and summaries.

Scope and evidence note. This report synthesises eight research notes compiled in October 2026. Figures labelled "vendor-reported" 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 [X](#) arXiv preprints were seen only through search summaries; they are marked where cited.

CONTENTS

Summary: main conclusions and limits

Speech acts and grounding define what is worth extracting

Machine pipelines work stage by stage, with uneven maturity

What shipped, what retreated, and how practitioners work

The solved, maturing and open map

A reference architecture: log, claims, beliefs, review

Ingested messages are an attack surface and a liability

Things you probably have not thought of

Navigating the field

Conclusion

About this report and its session log

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.

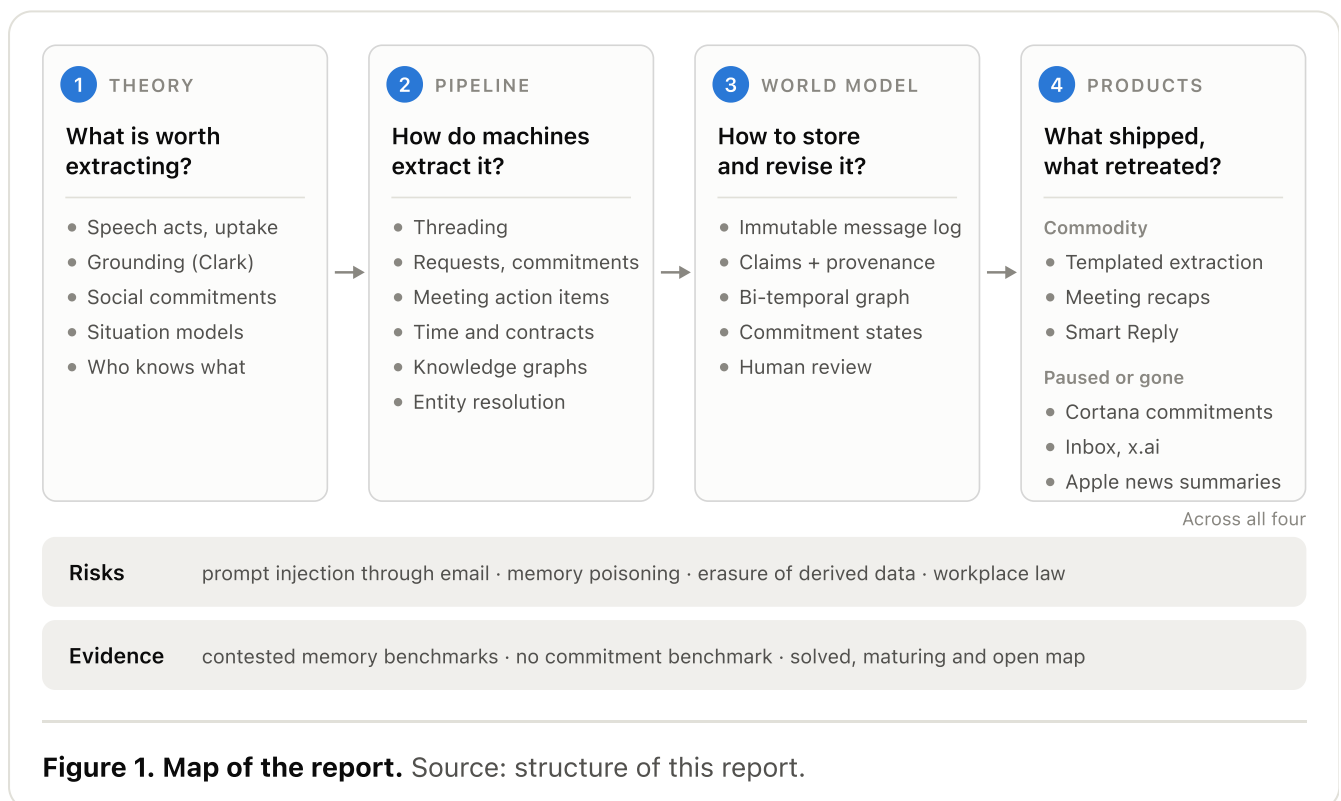
Main conclusions

- **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 reports 95-97% precision)
 - 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.
 -  Microsoft paused its commitments email (Briefing) in 2023 and gave no reason.
 - Resolving relative dates like "next Friday" and reasoning about durations are weak points for LLMs.
- **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, and one popular benchmark has 6.4% wrong answers. Only trust tests rerun on your own data.
- **Risks:**
 - Every incoming email is a way to inject instructions into the system. EchoLeak, a zero-click attack on 🇺🇸 Microsoft 365 Copilot, is the proven case.
 - Planted false memories can persist and affect later answers.
 - No regulator has said how the GDPR right to erasure applies to summaries and graphs derived from messages.

How the report is organised

Figure 1 is a map of the report: the four questions it answers in order, and the two concerns that cut across all of them. Figure 2 places the main papers, products and incidents on one timeline. Figure 3 takes eight everyday problems and sets how people cope next to how machines try, and where the machines are known to fail.



Foundations, 1962–1999

- 1962 ● **Austin**: utterances are actions, not only content
- 1980s ● **The Coordinator**: users tag each message as a speech act
- 1991 ● **Clark & Brennan**: grounding, the evidence of being understood
- 1991 ● **Wegner**: transactive memory, a “who knows what” directory
- 1993 ● **Suchman**: speech-act categories as discipline and control
- 1996 ● **Whittaker & Sidner**: the inbox doubles as a to-do list
- 1998 ● **Zwaan & Radvansky**: situation models indexed by time, space, entity, cause, intent
- 1999 ● **Singh**: social commitments, a debtor owes a creditor

Classic NLP and first products, 2004–2021

- 2004 ● **Cohen, Carvalho & Mitchell**: email speech acts; commitment F1 0.11–0.47
- 2010 ● **Lampert, Dale & Paris**: email zoning lifts request accuracy 72.3% → 83.8%
- 2013 ● **Gmail**: rule-based extraction from bookings and receipts in production
- 2016 ● **Smart Reply**: 10% of Inbox mobile replies
- 2017 ● **Cortana**: suggested reminders from promises in sent mail
- 2019 ● **Kummerfeld et al.**: 77,563-message chat disentanglement corpus
- 2019 ● **Azarbonyad, Sim & White**: commitment AUC falls 0.86 → 0.67 across companies
- 2019 ● **Inbox by Gmail** shut down
- 2020 ● **Google**: ML extraction reaches 95–97% precision
- 2021 ● **x.ai** acquired by Bizzabo

LLM era, 2023–2026

- 2023 ● **Microsoft** pauses the Briefing commitments email
- 2023 ● **Greshake et al.**: indirect prompt injection
- 2025 ● **Apple** pauses news notification summaries after false headlines
- 2025 ● **Zep vs Mem0**: the same benchmark scored 66.0 and 75.1
- 2025 ● **GraphRAG-Bench**: graphs win on reasoning, tie on lookups
- 2025 ● **EchoLeak**: zero-click injection through one email (CVSS 9.3)
- 2026 ● **Gmail AI Inbox**: suggested to-dos, US English first
- 2026 ● **EmailBench**: best agent passes 33.5% of scenarios

Figure 2. 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); Briefing email paused 2023
Reading indirect requests	Infer intent from politeness, hedges and context	Speech-act classifiers; LLM extraction	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	Durations 13.5–16% correct, off-by-one-day errors; no benchmark on message timestamps
Finding things again	Scroll and search; heavy folder users do no better	Threading, search, automatic 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 validated on one company (Enron) only
Summarising	Meeting notes and recap emails	LLM 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	Feels like surveillance: Suchman's critique, the Productivity Score backlash

Figure 3. 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**. Austin separated what an utterance says from what it *does* (asserting, promising, ordering). His felicity conditions add that a request or promise is fully constituted only through **uptake**: the addressee has to take it up, or the act "misfires" ([SEP Speech Acts](#)). Indirect speech acts ("could you maybe look at...") depend on inferring intention, and some theorists treat them as implicature rather than a fixed form ([SEP](#)). Grice's maxims explain why meaning so often exceeds the literal text ([Stanford handout](#)). For a builder, these two ideas set the problem's ceiling: surface patterns will miss polite, hedged and conditional requests, and a request is not a commitment until there is evidence it was accepted.

Winograd and Flores turned this into a product, The Coordinator. Users labelled each outgoing message with a conversation category, and 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. She argued that building systems on speech-act categories brings "**an agenda of discipline and control**" over the people categorised, and that 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 is to **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.

Multi-agent systems research gave commitments a data model. Singh argued that mentalistic (belief and intention) semantics fail in open systems and defined **social commitments**, which are directed (a debtor owes a creditor), arise in a context and can be manipulated in standard ways ([Singh IJCAI-99](#); [Singh AAAI-08](#)). Castelfranchi required explicit acceptance by the creditor; Singh considers that right only for some applications ([Commitments in MAS](#)). Yolum and Singh's **commitment machines** compile commitment-based protocols into finite-state machines and allow exceptional moves that fixed message sequences forbid ([Singh group](#)). Symboleo applies the same approach to legal contracts. It gives state machines for the lifetimes of contracts, obligations and powers, and handles subcontracting, assignment and novation ([Symboleo RE'20](#)). These are ready-made schemas, not just philosophy. 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.

Clark and Brennan's **grounding** model adds the epistemic side. Shared understanding is built step by step, each contribution needs evidence that it was understood "to a level sufficient for current purposes", and the medium changes how costly that evidence is ([Clark & Brennan 1991](#)). Email is asynchronous, so acknowledgement often never arrives. One message does not establish common ground. Discourse theories supply structure. Grosz and Sidner describe intentional and attentional structure, including a focus stack. RST and SDRT give rhetorical relations ([Grosz-Sidner discussion](#); [SDRT intro](#)). Together they suggest that a new message usually attaches to a recent open topic, which is a useful prior when linking messages.

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. Whittaker and Sidner found email doubling as a to-do list and archive, with filing failures causing messages to be "lost" ([CHI '96](#)). A ten-year follow-up across **600 mailboxes** found inbox sizes roughly unchanged while archives grew about tenfold ([Microsoft Research](#)). A field study of **345 users and more than 85,000 refinding actions** found heavy folder users were no more successful at refinding and were marginally slower. Search and threading helped, and scrolling was the most common behaviour ([bespacific summary](#); abstract-level only). Bellotti's Taskmaster grouped email, links and attachments into "thracks" (thread plus task). Users received it well, but it never became a full product ([CHI 2003](#)). Microsoft Research found that promises live in *outgoing* mail and that people "just rely on memory" or on flags to track them ([MSR blog](#)). An open-commitments ledger therefore replaces a fragile human function, and it is useful only if it can tell an open commitment from a discharged one.

Cognitive science suggests what dimensions a world model should track. Zwaan and Radvansky's event-indexing model says 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 ([Zwaan & Radvansky 1998](#)). Wegner's **transactive memory** shows groups spreading 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 ([Wegner, Erber & Raymond 1991](#)). The Pirolli-Card sensemaking model separates a foraging loop from a schema-building loop ([Wikipedia summary](#)). Its implication is that a machine world model should behave like a set of hypotheses revised as evidence arrives, not like a ledger of settled facts.

Machine pipelines work stage by stage, with uneven maturity

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

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**, which is framed as predicting reply-to links and then clustering. The reference benchmark is Kummerfeld et al.'s **77,563 annotated Ubuntu IRC messages**. That work also found that about 80% of conversations in the widely used Ubuntu dialogue corpus were missing messages or contained extra ones ([arXiv 1810.11118](#)). Getting whole threads exactly right is still hard. On that test set, supervised models reach **cluster F1 of 36.6-51.7**. A 2026 workshop paper reports prompted GPT-4.1 at 47.5 and Gemini-2.5-pro at **61.3**; it has not been independently replicated ([CEUR Vol-4178](#)). A separate LREC 2026 study found out-of-the-box GPT-4o "performs poorly", while a fine-tuned GPT-4o Mini reached 74.4 link F1 but only 45.3 cluster F1 ([ACL Anthology 2026.lrec-1.229](#)). Zero-shot ChatGPT and GPT-4 trailed supervised BERT on IRC speaker identification by 25.6 and 12.3 points ([arXiv 2310.16301](#)). The practical reading is to use platform-native thread IDs as primary keys for *structure*, and to treat any content-inferred thread as a revisable cluster. Threads are also not topics. A 2026 Slack case study, a preprint seen only via search, argues that one thread can cover several unrelated topics ([arXiv 2605.19627](#)). The notes found no benchmark for linking one topic across threads.

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

Classifying email by intent is a 20-year-old, small-data field. Cohen, Carvalho and Mitchell's verb-noun taxonomy (Request, Propose, Commit, Deliver $\times$ Meeting, Information) reached **precision above 80% and recall above 50% for some classes** ([EMNLP 2004](#)). In the full paper, Commitment F1 ranged only 0.11-0.47, and annotator kappa sat between 0.72 and 0.83 ([PDF](#)). Lampert, Dale and Paris showed that preprocessing matters more than the model. On Enron, a request classifier rose from **72.3% to 83.8% accuracy once email "zoning" separated quoted, forwarded and signature text** (a 41%+ error reduction). Human agreement was only kappa 0.681, and conditional or indirect requests drove the disagreement ([NAACL 2010](#)). 🇺🇸 Microsoft's sentence-level commitment work reached **F1 0.78-0.81 within one corpus**. Moved from Avocado to Enron, F1 fell to 0.73 and **AUC fell from 0.86 to 0.67**; adversarial domain adaptation recovered part of the loss ([Azarbondy, Sim & White, WSDM 2019](#)). Smart To-Do reported 86% precision and 84% recall for commitment sentences on Avocado, against an annotator kappa of 0.694 (✗ [arXiv 2005.06282](#)). 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, a September 2026 preprint seen via search, in which the best of eight agent configurations passed **33.5% of 206 scenarios** even though 99.7% of its tool calls succeeded ([arXiv 2609.31906](#)).

Meetings: summaries are measurable, action items barely are

Action-item detection has very little labelled data. AMI has no direct action-item labels. Its 381 items across 101 meetings are derived by linking dialogue acts to summary sentences, and ICSI's annotations are no longer public ([Liu et al., ICASSP 2023](#)). 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. On the QMSum Mistake set (200 LLM summaries, 9 error types), GPT-4 Turbo detected errors with about 89% average accuracy but only 72% for hallucination (✗ [arXiv 2407.11919](#)). Structured pipelines (FRAME) cut hallucination and omission scores by 2 of 5 points ([arXiv 2509.15901](#)). One signal matters for owner extraction. In Meeting Delegate, phonetically corrupted names, used as a stand-in for ASR errors, cut GPT-4o's explicit-cue response rate from **94.3% to 68%** ([arXiv 2502.04376](#)). That paper did not measure action items, so the effect on assignee extraction is inferred. The notes found **no verified LLM action-item F1 on AMI or ICSI**.

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

Where the schema is fixed, conversation-to-state is close to solved. Cleaning MultiWOZ's labels (version 2.4) raised every state-of-the-art tracker's score (✗ [arXiv 2104.00773](#)). Correcting inconsistent tags added 7-10% joint goal accuracy, while unseen entities cut it by 29%, a sign of entity memorisation ([Qian et al., SIGDIAL 2021](#)). Open relation extraction from dialogue is harder. On DialogRE, BERT scored 61.2 F1 overall but 55.4 when relations had to be found early in a conversation (📄 [Papers with Code](#)). LLMs are competitive there but weak on inverse relations ([IJCAI 2024](#)).

Time is a specific weak spot. In the synthetic Test of Time benchmark, frontier models scored about 90-93% on "when did X happen" lookups. On timelines they scored 24-36%, and on duration arithmetic **13.5-16%**, with frequent off-by-one-day errors ([ICLR 2025](#)). 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**. The design inference is to let the LLM output the unresolved expression and to do the date arithmetic deterministically.

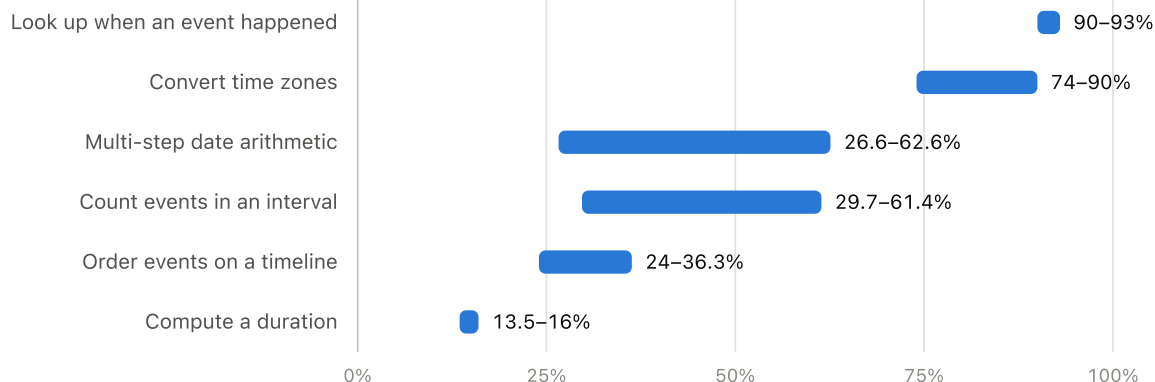


Figure 4. LLM accuracy on time questions. Source: [Test of Time, ICLR 2025](#). Each bar spans the scores of GPT-4, ♦ Gemini 1.5 Pro and ✱ Claude 3 Sonnet (2024 models); the lookup row is approximate.

Contract clause extraction from formal documents works at "**junior legal assistant**" level. ContractEval ran 19 LLMs on CUAD (510 contracts, 41 clause types). Open models over-answered "no related clause", and thinking mode lowered correctness (✗ [arXiv 2508.03080](#)). The best configuration in another CUAD study reached recall 0.77 and precision 0.66 ([arXiv 2505.12837](#)). The notes found no benchmark for obligation extraction from informal dialogue. The nearest relative, CraigslistBargain, captures only the agreed price ([arXiv 1808.09637](#)).

Graph construction, entity resolution and memory systems

LLM knowledge-graph construction comes in two schools. Schema-based extraction is precise but scales poorly; schema-free extraction is flexible but inconsistent ([survey](#) ✗ [arXiv 2510.20345](#)). KGGen's author-run MINE benchmark scores fact capture at **66% for KGGen**, **48% for GraphRAG** and **30% for OpenIE**, with backbone models ranging from 44% to 73%. The authors acknowledge that "over or under de-duplication" remains, and their test texts are encyclopedic articles, not conversations ([arXiv 2502.09956](#)). For person and organisation identity, classical Fellegi-Sunter matching is mature: Splink trains unsupervised and scales past 100M records ([PyPI](#)). Generative LLMs match or beat fine-tuned models with little training data and generalise better to unseen entities ([Peeters et al., EDBT 2025](#)). Senzing's "principle-based" resolution reasons about how attributes behave, for example shared values versus stable ones. These are vendor claims ([Senzing](#)). The notes found **no measured duplicate or merge-error rate on conversational data**.



Schema-first

Define the types, then extract into them

EXAMPLES

Palantir Ontology · Gmail templated extraction · MultiWOZ slots · Symboleo contracts

STRENGTHS

- + Precise, consistent types
- + Easy to query, check and audit
- + Proven at scale on templated mail

WEAKNESSES

- Misses whatever the schema does not name
- New types need modelling work
- Brittle on entities it has not seen

RESULTS

- Gmail ML extraction: 95–97% precision
- Dialogue state: fixing labels adds 7–10% accuracy
- Unseen entities cut that accuracy by 29%



Schema-free

Let the model propose entities and relations

EXAMPLES

OpenIE · KGGen · GraphRAG · LightRAG

STRENGTHS

- + No upfront modelling
- + Captures the long tail
- + Helps multi-hop and summary questions

WEAKNESSES

- Inconsistent types and relations
- Duplicate or wrongly merged entities
- High token cost per query

RESULTS

- Fact capture: KGGen 66%, GraphRAG 48%, OpenIE 30% (author-run)
- Same method, different model: 44% to 73%
- Reasoning 53.4 vs plain RAG 42.9; lookups tie
- Up to ~100K prompt tokens per query (LightRAG)
- Mem0 removed graph memory from its open-source SDK



A hybrid · this report's synthesis

A schema for the core (people, commitments, events, time); open extraction for the long tail, promoted into the schema once a type keeps recurring; every claim keeps its source.

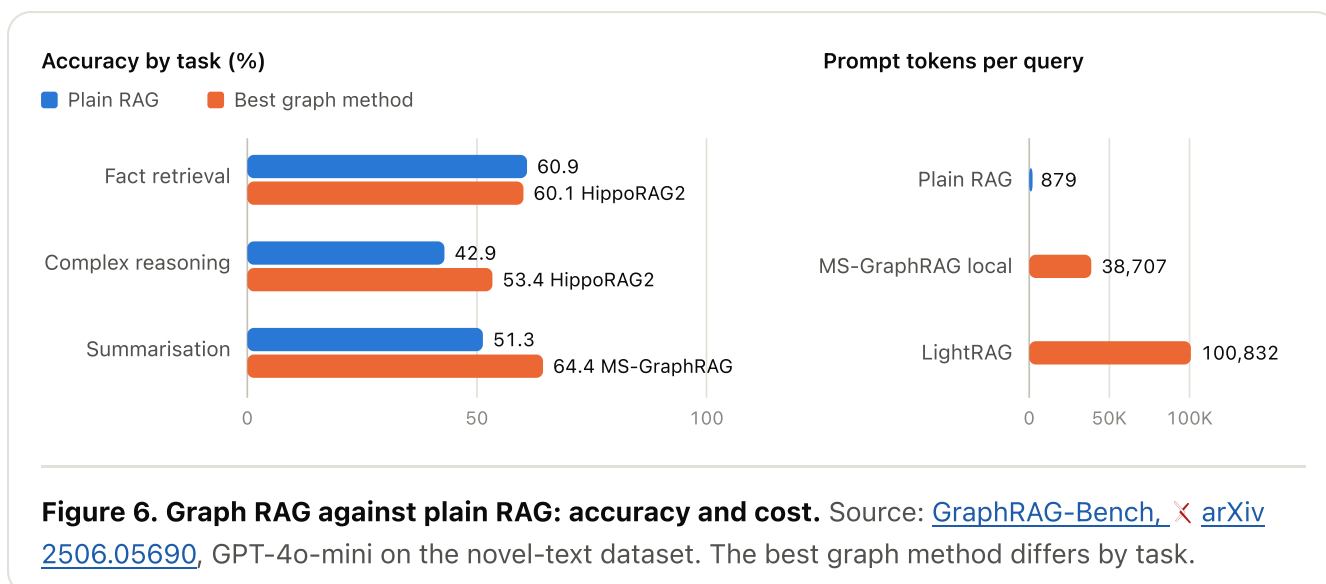


Open questions · no answers found

How should a schema grow from open extractions? What are real duplicate and merge-error rates on conversations? Does a graph still win against million-token context? Who approves a new type?

Figure 5. Two schools of graph construction. Sources: [survey](#), [arXiv 2510.20345](#), [KGGen](#), [GraphRAG-Bench](#), [Qian et al.](#), [CIDR 2020](#), [Mem0 docs](#). The hybrid is this report's synthesis.

Agent-memory products fall into three families. Graph-indexed RAG includes GraphRAG, LightRAG and HippoRAG 2. Conversational fact stores include Mem0, Zep/Graphiti and MIRIX. Agent-managed files include Letta and 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. In GraphRAG-Bench, plain RAG scored 60.9 on fact retrieval against HippoRAG2's 60.1. On complex reasoning HippoRAG2 led 53.4 to 42.9, and on summarisation MS-GraphRAG led 64.4 to 51.3. Prompt tokens per query ranged from about 880 for vanilla RAG to 38,700 for MS-GraphRAG local search and 100,800 for LightRAG ([arXiv 2506.05690](#)). 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.



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.**

G Google's **M** Gmail extraction (Juicer) is the most rigorous production evidence and covers *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, and more than 45,000 lines of rules were deleted ([CIDR 2020](#)). The system serves over a billion users without engineers seeing individual emails ([Google Research](#)). Senders can also supply structure themselves through schema.org markup ([Google developers](#)). Smart Reply shows that tightly constrained generation scales: it produced 10% of mobile replies in Inbox ([KDD 2016](#)). **None of this is evidence for extracting human commitments.**

M Microsoft shipped commitment detection and then retreated. Cortana's 2017 "Suggested Reminders" spotted promises in sent mail ([eWeek](#)). The Briefing email listing outstanding commitments was **paused on 15 January 2023**, with no stated reason ([Microsoft support](#)). Cortana itself reached end of support, and its email feature Play My Emails was retired in June 2024 ([Microsoft Support](#)). **G** Google Inbox closed in 2019 as part of product consolidation. Snooze, nudges and **M** Smart Reply moved into Gmail, but bundles did not ([Wikipedia](#)). The assistant-as-participant model did not survive as a standalone business: x.ai was acquired by Bizzabo in June 2021 ([CB Insights](#)), and Clara Labs' acquisition dates conflict across sources. **A** Apple's notification summaries falsely reported a news event in December 2024, so Apple paused news summaries in iOS 18.3. It restored them in the iOS 26 beta with the warning "Summarization may change the meaning of the original headlines" ([AppleInsider](#)). The fix was a UI disclaimer, not better accuracy. Gmail's January 2026 "AI Inbox" added suggested to-dos, starting with US English users (secondary sources only) ([TechRepublic](#)).

Meeting tools are commodity, and their accuracy is not measured. A university review of ^{zoom} Zoom AI Companion found action items accurate "**when explicitly stated**" ([Seattle U](#)). An Otter pilot found them "sometimes incorrect" ([Univ. of Alabama](#)). A CSCW 2025 study of an LLM recap prototype with seven  Microsoft employees found users adding, editing and deleting content, and noted hallucination risk on long dialogue ( [arXiv 2307.15793](#)). **No vendor publishes action-item precision or recall**, and the notes found no independent product study. Vendor-reported impact numbers do exist. Microsoft's randomised Copilot trial (66 firms, 7,137 workers) reported that regular users cut email time by about 3.6 hours a week. The intent-to-treat effect was about 1.3 hours, and only about 40% of those offered access used it regularly ([arXiv 2504.11436](#), vendor-authored).

Two older industries already solved parts of the problem. **eDiscovery** treats a thread as a graph with "inclusive" terminal messages that contain all earlier text, so reviewers read only the minimal set.  Microsoft Purview flags both branches of a fork as inclusive, where Outlook's conversation view could hide unique mid-thread content ([Purview docs](#); [Relativity](#)). Enterprise graphs take two approaches.  Palantir's Ontology is schema-first: humans define objects, links and actions, and data is mapped in rather than learned from text ([Palantir docs](#)).  Atlassian's Teamwork Graph builds from tool metadata and permissions, and opened through an MCP server in May 2026 ([Nasdaq release](#)). Contract-management vendors publish no comparable extraction benchmarks, and 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

Figure 7 names the problem domains of the field and marks each one; the table below gives the evidence for each mark.

24 named problem domains

5 solved

9 maturing

10 open

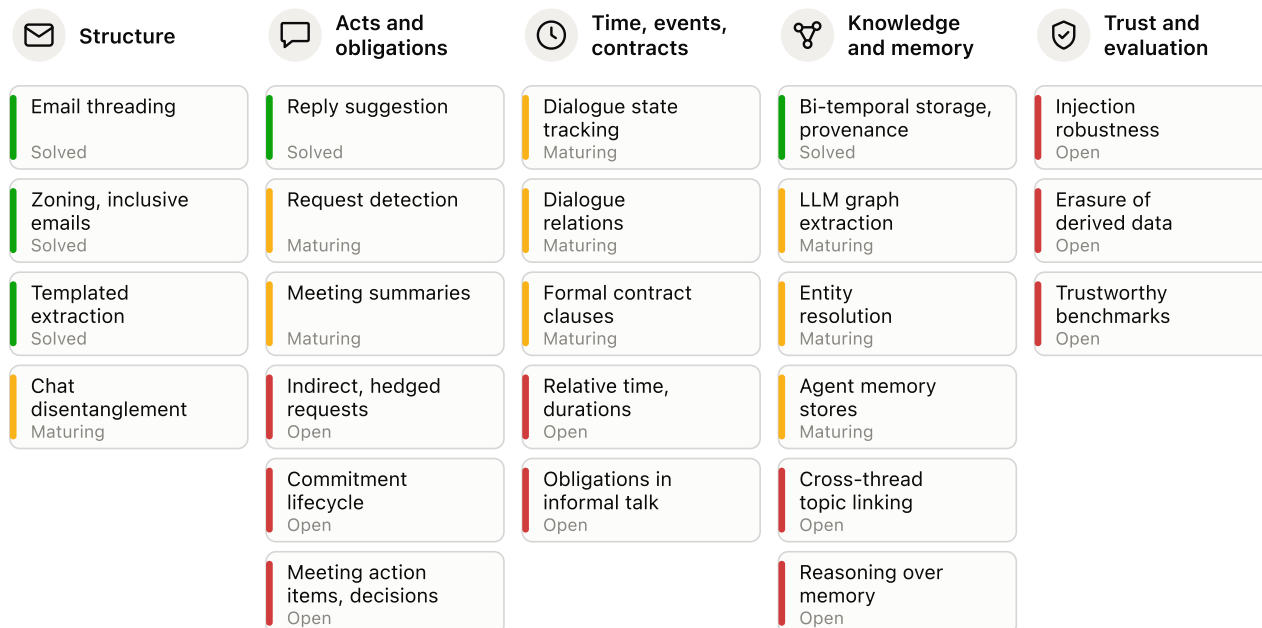


Figure 7. The problem domains and their status. Source: the table below. A domain is "open" when no method or benchmark shows it working on real conversations.

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-O ; Splink
● Solved (research)	Email zoning as preprocessing	+11.5 accuracy points on requests (Lampert 2010); zoners brittle across corpora

Status	Capability	Evidence and caveat
● 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 ~100x 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.

Figure 8 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.

19 sub-problems behind one goal

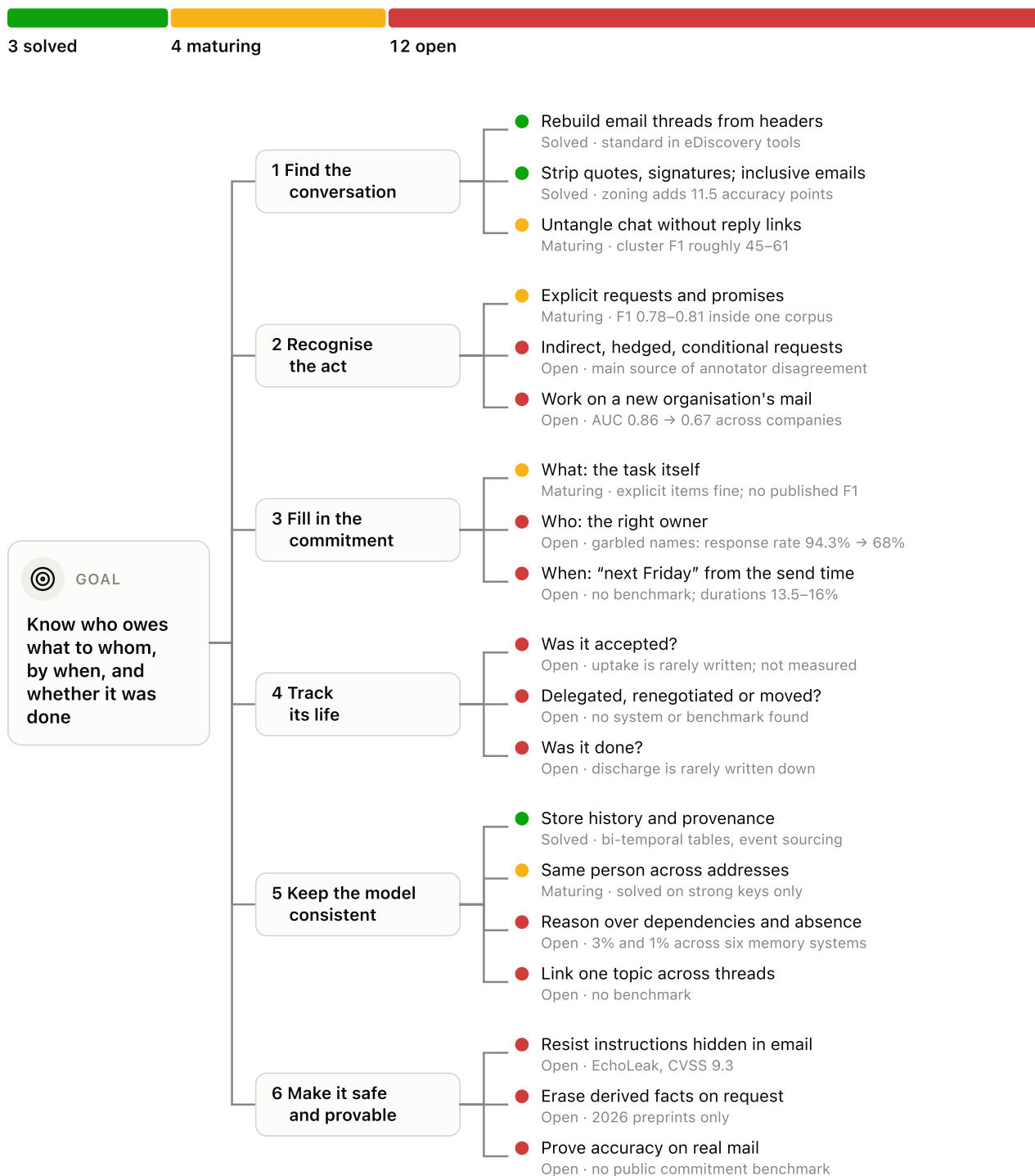
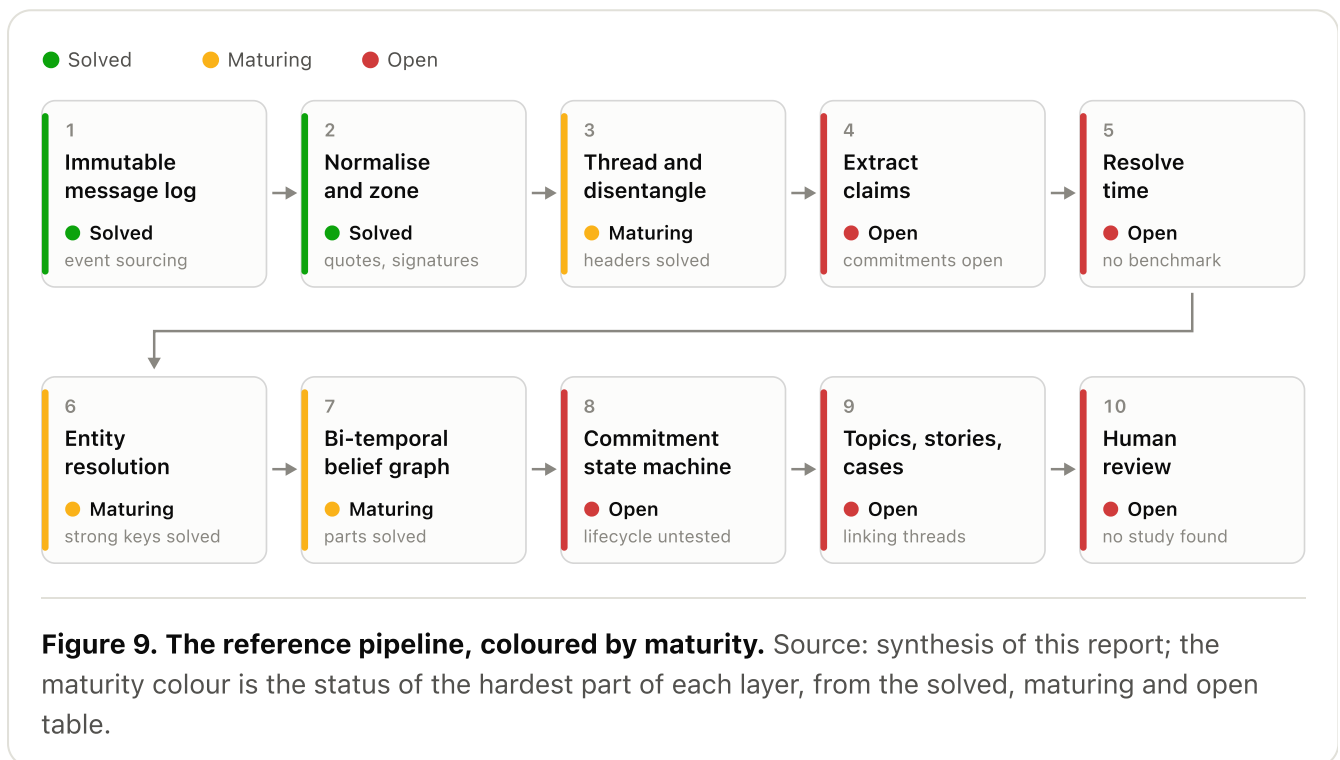


Figure 8. The problem tree behind "who owes what". Source: synthesis of this report; each leaf's evidence is cited in the sections above.

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

Layer	What it does	Grounding	Status
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
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

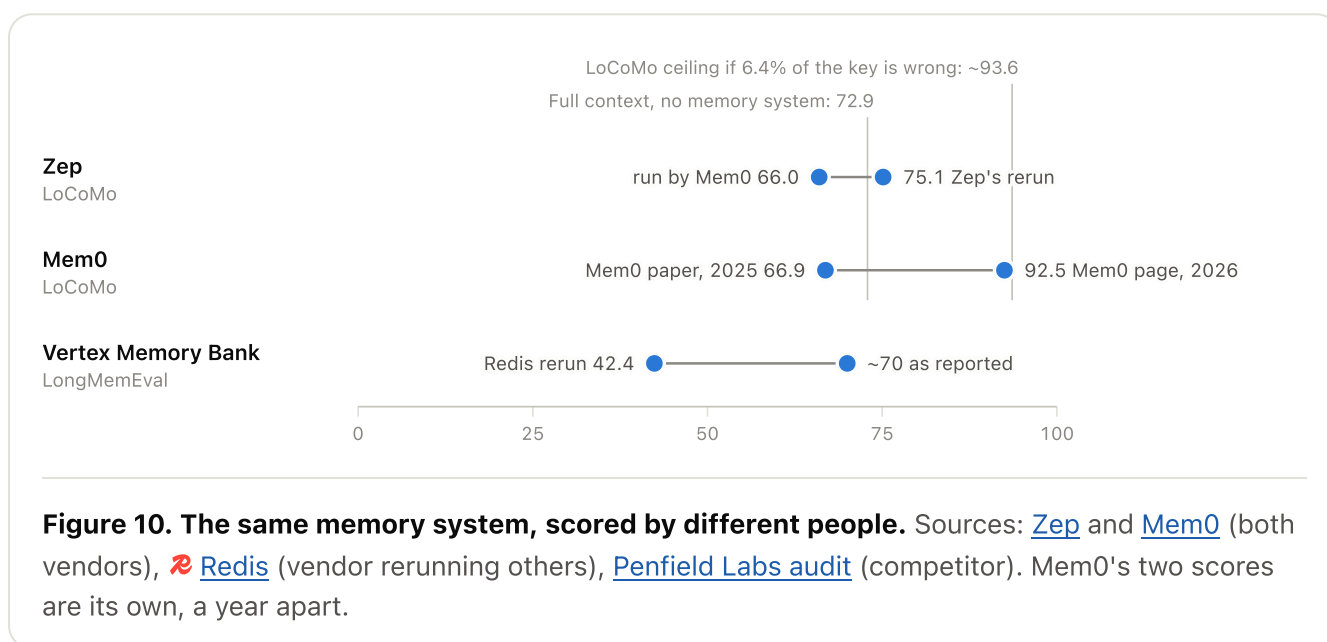
A request is not a commitment, and silence is not acceptance. Speech-act uptake and Castelfranchi's acceptance condition mean every directive needs a state of proposed, accepted, declined or unanswered. Clark's grounding means "sounds good" may not cover everything in the thread. Discharge is rarely 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 [G 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](#)).



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 ([Azarbondyad](#)). 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 [G 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

Datasets	What it gives you	Caveat
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](#), [Azarbonyad 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.

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.