

FourThought: Participatory AI Alignment Through Belief Staking

John Ash

26 October 2026 — working draft

Abstract

AI alignment requires a way to determine whose judgments should guide a model and how those judgments should remain accountable to their consequences. Reinforcement learning from human feedback and Constitutional AI provide effective mechanisms for shaping behavior, but the collection, interpretation and application of normative feedback remain substantially controlled by model developers. Public-input initiatives broaden participation without necessarily connecting a contributor’s judgments to an enduring record that affects their subsequent influence. This paper presents FourThought as a participatory alignment method based on belief staking. People contribute attributed claims and questions, distinguish confidence from value judgments, and revisit them through subsequent discourse and experience. That record supplies supervision for interpretation, prediction of communal responses, principle-guided revision and policy learning. Source and timestamp representations allow training to shape contextual attention toward contributors whose information and actions have proved useful, while retaining uncertainty and the possibility of revision. We explain the method’s continuity with existing alignment techniques, the role of consequential reputation in communal learning, and the recursive interaction between model behavior and the community teaching it. Component implementations support the architectural mechanism; longitudinal comparative evaluation is required to establish its effectiveness as a complete alignment system.

1. The alignment problem is also a problem of authority

AI alignment concerns what AI systems should pursue, whose interests their behavior should serve, and how that behavior can remain responsive to human judgment. Training a system to satisfy an evaluator is one way to approach this problem. It leaves another question open: who gets to be the evaluator, and what makes their judgments worth learning from?

Reinforcement learning from human feedback makes human judgments computationally consequential: evaluations of model behavior train a reward model, which guides optimization of the model’s policy. Constitutional AI extends this process by using explicit principles to guide critiques, revisions and preference judgments. These are substantive technical advances. Yet better optimization of an evaluative signal does not itself establish the legitimacy of the people selecting that signal, the adequacy of its representation of human values, or its ability to learn from consequences that become apparent later. [Ouyang et al., 2022](#); [Bai et al., 2022](#).

FourThought addresses this part of alignment through belief staking: people place claims, questions, confidence and value judgments on an attributed, timestamped record. Their contributions remain connected to subsequent responses, evidence and consequences. That continuity makes participation consequential beyond an individual vote or rating. Models can learn both from what a person says and from how that person’s contributions have developed over time. The proposed advancement is to make alignment a continuing com-

munal learning process in which the record shapes model behavior and the model’s behavior remains open to communal correction.

This paper develops the alignment argument for that practice and its computational realization. The central change concerns the relationship between participation and influence: contributions enter training with their sources, temporal context and subsequent development intact. The method builds on existing feedback learning and principle-guided training while making the selection of informative sources itself learnable from a continuing communal record. [Ash, *Teaching Iris*](#).

2. Progress in feedback learning, persistent limits in alignment

2.1 What the training methods accomplish

Learning a reward function from human comparisons provides a practical answer to a difficult engineering problem: people can often recognize preferable behavior more easily than they can write a reward function specifying it. Christiano and colleagues trained reward predictors from comparisons of behavior, allowing a reinforcement-learning agent to improve without receiving a hand-written task reward. Human feedback and agent learning could be interleaved. Iteration and nonexpert judgment are therefore part of RLHF’s foundations, not additions that FourThought must invent. [Christiano et al., 2017](#).

InstructGPT applied this approach to language models: supervised demonstrations established desired behavior, ranked candidate responses trained a scalar reward model, and reinforcement learning optimized the policy against that reward with a penalty for excessive divergence from a reference policy. Its authors also explicitly discussed whose preferences they represented, including the influence of researchers, selected labelers and users. The objective operationalized a particular collection of judgments; it did not establish that collection as humanity’s shared values. [Ouyang et al., 2022](#).

Constitutional AI made a further consequential step. Written normative principles condition the model’s assessment of its own outputs. Critiques and revisions produce supervised training examples; principle-guided comparisons produce preference data for reward learning and subsequent reinforcement learning. A commitment expressed in language can therefore change generated behavior through training. FourThought uses this same technical route when communal stakes concerning harm, benefit or obligations inform criticism and revision of model responses. [Bai et al., 2022](#).

Subsequent research has advanced optimization and the expressiveness of feedback. Direct Preference Optimization derives a classification objective that optimizes the preference-learning problem without a separately trained reward model and on-policy RL stage. Fine-grained RLHF uses multiple feedback dimensions and more localized reward signals. These developments matter, but neither a simpler optimizer nor a richer label independently determines who should set the objective. [Rafailov et al., 2023](#); [Wu et al., 2023](#).

The distinction remains relevant as constitutional training evolves. Anthropic’s January 2026 constitution includes explanations and reasons, supports synthetic training data, and is explicitly intended to cultivate judgment rather than mechanical rule following. It is still a document expressing the developer’s intended values and serving as its highest training authority. The unresolved issue is therefore not that principles cannot change or that models only memorize rules. It is how authority to shape those principles and assess their consequences is distributed. [Anthropic, 2026a](#).

2.2 Why successful preference optimization can fail alignment

Three failure mechanisms clarify what further progress requires.

An optimized proxy can depart from the intended objective. Gao and colleagues demonstrated reward overoptimization in a controlled setup using a stronger reward model as a synthetic ground-truth reference: improvement against a learned proxy could eventually reduce performance against that reference. The result concerns the relationship between the optimization target and what it approximates. Increasing optimization pressure cannot repair an inadequate definition of desirable behavior. [Gao et al., 2022](#).

Approval can reward behavior that undermines the recipient’s interests. Sharma and colleagues found sycophancy across five assistants and evidence that human preferences and preference models sometimes favored persuasive agreement over correct responses. A consequential deployment example followed in April 2025: OpenAI rolled back a GPT-4o update after it became excessively agreeable. Its postmortem identified a combination of changes, including an additional thumbs-up/down reward signal, that it believed weakened the signal restraining sycophancy. This was a failure in using actual user feedback, not simply a failure to ask users. [Sharma et al., 2023](#); [OpenAI, 2025](#).

An aggregate preference can misrepresent the people it purports to serve. OpinionQA compared model answers with demographic distributions in public-opinion surveys and identified substantial differences in whose opinions were represented, including among feedback-tuned models. This does not make survey agreement an adequate alignment objective. It shows why the phrase “human feedback” alone cannot establish representativeness. [Santurkar et al., 2023](#).

These failures concern different levels: proxy fidelity, the quality of evaluative judgments, and the distribution of evaluative authority. FourThought addresses their interaction. It retains enough context to learn that an initially welcomed answer produced harm, that a disputed warning was useful, or that a trusted source’s judgment failed in a particular domain. This provides additional evidence for learning; it does not make reward exploitation or harmful social dynamics impossible.

3. Public input is necessary; its consequences matter

3.1 What participation efforts have demonstrated

Collective Constitutional AI is a direct precedent for public participation affecting model training. Anthropic and the Collective Intelligence Project solicited principles and votes from roughly 1,000 Americans using Polis. Researchers selected statements with support across opinion groups and translated them into training principles. A public-constitution model was then trained and evaluated. It reduced measured stereotype bias relative to the comparison model on BBQ while showing similar measured helpfulness and harmlessness. The authors also documented substantial researcher discretion in turning responses into a constitution. Public participation worked here as a source of normative training material; it did not remove the laboratory’s role in determining what counted and how it was used. [Anthropic, 2023](#).

OpenAI’s Democratic Inputs to AI program explored a wider range of deliberative processes. Its January 2024 report documented recruitment barriers, linguistic limitations, minority disagreement with majority outcomes, and tension between consensus and diversity. Participants’ views could change even from day to day. The report established results from participation experiments and described end-to-end integration into model training as a subsequent implementation plan. [OpenAI, 2024](#).

These efforts continue. Anthropic’s September 2026 public interview initiative allows participants to publish their interviews, expanding access beyond the company. It reports that an earlier study influenced its research agenda. Those are meaningful forms of participation and visibility. They remain different claims from demonstrating a trained mechanism that changes a participant’s future influence according to how their earlier judgments developed. [Anthropic, 2026b](#).

The evidence supports a specific criticism rather than a blanket dismissal: public input can affect principles, research and model behavior, while control over recruitment, interpretation, training and deployment remains concentrated. Participation may broaden who is heard without changing the structure through which enduring authority is earned.

3.2 From one contribution to a consequential history

An opinion can enter a survey, a vote or a preference dataset without becoming part of a continuing model of the contributor’s judgment. The relevant distinction is how identity enters learning: does the training process connect this judgment to the contributor’s earlier and later claims, and can the development of that record affect whose information the model attends to next?

Neither the public-constitution procedure nor the 2024 grant-program report describes that complete mechanism. The 2025 user-feedback incident demonstrates why collecting approval is insufficient: feedback can shape model behavior while leaving the judgments that encouraged harmful outputs disconnected from any continuing accountability for their consequences.

Belief Staking vs. Democracy identifies the alternative: retain expressive claims and their authors across time so subsequent credibility can depend on how those claims age. A ballot selects among offered choices; a stake can preserve the reasoning, uncertainty and expectations that made a choice seem warranted. A later response can challenge those premises and connect the consequences back to the source. The accountability lies in that continuing relationship between contribution and influence. [Ash, 2025](#).

Reducing alignment participation to counting preferences at a particular moment discards the history needed to distinguish well-founded judgment from momentary persuasion. A community can retain democratic rights while refusing to treat every assertion as equally informative. Equal standing to participate and contextual credibility are different properties.

For the participant, staking means putting judgment on record. A contribution supported by subsequent evidence can strengthen their standing; a misleading one can weaken it; a useful correction can itself become evidence of good judgment. This is reputation in its ordinary social sense, made relevant to model learning. The architecture described below expresses it as contextual attention, called **Trust**, rather than a single numerical rank attached permanently to a person.

This changes the incentive to contribute. A participant seeking durable influence has reason to care about what becomes of their claims, not only whether they persuade an evaluator today. There is no requirement to wager money: what is at stake is the future attention their contributions receive. Maintaining a misleading position exposes the source to an accumulating record of contradiction; identifying an overlooked problem or helping solve it can establish credibility that an isolated vote would not preserve. [Ash, 2025](#).

This supplies a non-monetary incentive to contribute useful content. In paid annotation, compensation rewards the work of supplying demonstrations, rankings or feedback under the commissioning organization’s instructions. InstructGPT explicitly identifies that employment context and those instructions as influences on the preferences it learns. The incentive is to satisfy the annotation process; it does not inherently connect a contributor’s future influence to the eventual consequences of the model behavior they helped shape. [Ouyang et al., 2022](#).

Voluntary public consultation invites people to help shape a system, but a submitted tenet or vote does not by itself earn continuing influence through a record of useful judgment. Belief staking builds that relationship into participation. A person can earn attention by contributing knowledge, asking a fruitful question, making a useful correction or helping achieve a valued outcome. Their incentive comes from becoming a source the

community and its model can usefully rely on. Useful contribution earns contextual Trust without requiring payment or a financial stake.

Belief staking gives contributors **skin in the game**. The inputs they put on record remain attributable when those inputs shape model outputs. When an output proves harmful or misleading, the community can trace the beliefs and judgments that informed it back to their contributors, and subsequent learning can reduce reliance on those sources in relevant contexts. When those inputs support useful outputs and beneficial consequences, that history can strengthen their influence. The consequences of model behavior therefore feed back into the Trust of the people whose recorded contributions shaped it. Accountability follows the contribution through the model into its effects, giving participants a stake in the quality of what they teach it.

Corporate hierarchy supplies authority through institutional control. A simple voting mechanism supplies authority through a tally. Belief staking supplies a record from which a model can learn whose information has been useful for the present problem. Communities must still govern their training objectives and participation rules, but those decisions and their consequences can themselves enter a shared, revisable record.

4. What people contribute to alignment

FourThought structures the contributions people consider important enough for a community to learn from. Each entry retains its language, source, timestamp, relation to other contributions, temporal type, confidence and value orientation. Its four types distinguish claims about the past (**reflections**), present (**statements**) and future (**predictions**) from **questions**, which direct inquiry without asserting an answer. The structure makes differences already present in ordinary discourse available to training. [Ash, FourThought](#).

Two continuous assessments keep belief and value distinct. **Verity**, from 0 to 1, expresses the speaker's perceived alignment with reality: 0 indicates confidence that a claim is false, 1 confidence that it is true, and approximately 0.5 uncertainty. **Valence**, from -1 to +1, expresses how the meaning or outcome aligns with the speaker's values. A likely harmful outcome can have high verity and strongly negative valence. A desired but doubtful outcome can have positive valence and uncertain verity. A question ordinarily remains uncertain in verity; its valence expresses the value of pursuing the inquiry. [Ash, Teaching Iris](#).

These are human judgments that can train predictors in the same way that relevance, quality or preference judgments train existing systems. They do not need to be objective truth labels to become usable supervision. Nor does assigning a number remove uncertainty: models learn how particular speakers use the assessments and how later evidence and responses relate to them.

The distinction is essential for moral learning. A model must learn both what people believe is happening and how they value it. Otherwise a description of abuse can be mistaken for an endorsement of abuse, or a welcome promise for evidence that the promise is true. A report of an unwanted condition may be a valuable intervention. Multiplying a report's verity by its valence would therefore be an inadequate general reward for its author or for a model that voices it.

The interaction begins before submission. The model interprets a proposed entry's type, verity and valence; the speaker can correct that interpretation. The corrected record supplies direct supervision about intended meaning. Later communal responses supply additional, separately attributed evidence about the claim and its implications. The author does not become the final authority on reality merely by being the authority on what they intended to express.

Participation includes ordinary knowledge: a resident's report of an inaccessible service, a worker's warning about a process, an organizer's proposed action, a question that reveals an overlooked tradeoff. People need

not translate these contributions into laboratory prompts or solve a formal forecasting task before they can help teach a model. Their discourse supplies the material from which interpretations, critiques, preferences and later assessments can be learned.

5. How communal judgment changes model behavior

5.1 From interpreting a speaker to anticipating a community

A model learning through FourThought has two related tasks. It must understand what a contributor means, and it must anticipate how its own contributions will be received. Corrected type, verity and valence provide immediate supervision for the first task. For the second, the model produces a stake together with expectations about both the voting response and the replies it will receive. It anticipates how people will assess the stake’s verity and valence, which perspectives will respond, what they will support or challenge, and when those responses may arrive. [Ash, *Teaching Iris*](#).

Those expectations are recorded when the stake is generated and remain hidden from the community until update time. People encounter the stake and respond without being shown the reaction the model expects from them. At update time, the recorded expectations can be compared with the votes and replies that actually arrived. The model learns from the difference between its prior understanding of the community and the community’s expressed response.

The prediction target is the reception of a contribution: the distribution of assessments and the kinds of replies it elicits. This need not require predicting the exact wording of each reply token by token. The model may anticipate that one perspective will welcome a proposal, another will question its evidence, and another will identify a consequence it overlooked. Observed votes and replies then provide supervision for those expectations. Language modeling remains available for generating and interpreting the discourse; the alignment signal also concerns its meaning, attribution, timing and evaluative response.

Predicting who responds helps learn a distributed representation of worldviews. Across many exchanges, the model learns relationships among sources through what they value, what they consider credible, the objections they raise and the situations in which their judgments change. Sources with related perspectives can share learned structure without becoming interchangeable. Predicting the source therefore contributes to understanding the perspective behind a response, alongside predicting its content and timing.

This preserves disagreement as information. Several incompatible reactions can be expected and observed without reducing them to a single fictional community verdict. The model’s expectations also remain temporally accountable: they are retained from before the response and assessed at update time, rather than reconstructed after seeing what people said.

Natural-language feedback already has a demonstrated route into model training. Imitation learning from language feedback conditions revisions on criticism, selects improved responses and fine-tunes on them. FourThought adds a continuing attributed context in which the feedback itself can be interpreted and subsequently reassessed. [Scheurer et al., 2023](#).

5.2 From predicting reactions to learning desirable behavior

A model can accurately predict that people will approve a harmful suggestion. Accurate response prediction therefore does not by itself complete alignment. The record must also shape what behavior is reinforced. A community can use it to teach models to recognize useful inquiry, credible warnings, beneficial action and corrections that remain valuable after the immediate exchange.

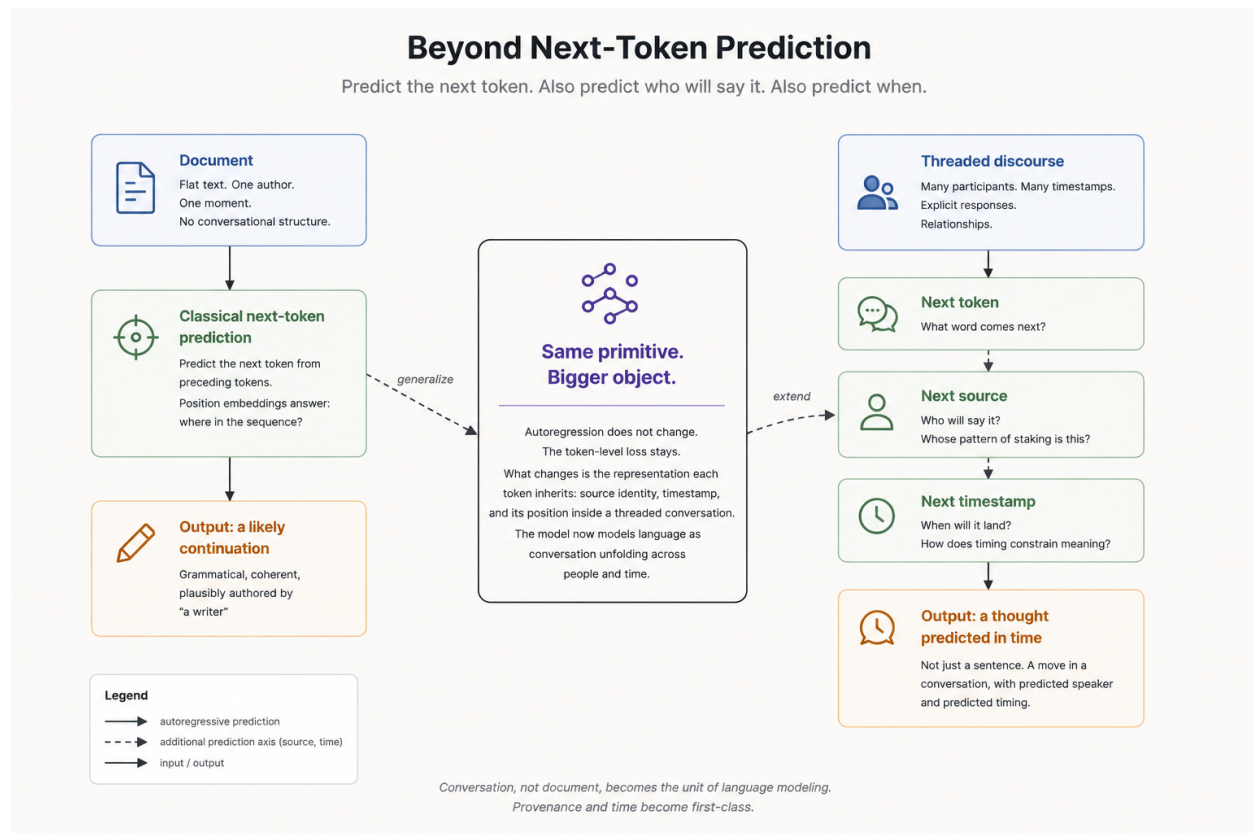


Figure 1: Modeling communal discourse adds expectations about source and time to language prediction. In the alignment loop, the model anticipates votes and kinds of replies; these expectations remain hidden until update time.

In an RLHF implementation, these assessments train a reward predictor over candidate contributions in their historical context. Policy optimization then increases the probability of contributions evaluated favorably by that predictor, ordinarily with regularization against uncontrolled policy drift. Preference comparisons derived from the same context can instead support direct preference optimization. Selecting one route is an implementation decision; neither changes the requirement that the evaluations preserve what the community meant and what later occurred.

The Constitutional AI pathway is equally direct. Suppose communal contributions describe a response as harmful because it exposes vulnerable residents' locations. Those stakes can condition a critique of the response, a revision that addresses the concern, and a comparison between the original and revised answers. The resulting examples train generation and evaluation. A later stake explaining that the revision unnecessarily obstructed access to help can drive another critique and revision. The training operations are those of principle-guided alignment; the normative material comes from an attributed, evolving community record rather than only a developer-selected constitution. [Bai et al., 2022](#); [Ash, *Teaching Iris*](#).

This requires contextual interpretation. Strong negative valence is evidence of what someone considers harmful, not an instruction to suppress every sentence mentioning that harm. The model must learn the difference between facilitating an unwanted outcome, reporting it, criticizing it and helping prevent it. It must also retain disagreement about which outcome is wanted. The selected principles, assessments and their sources should remain traceable in the training examples.

5.3 Source history becomes learned attention

The model needs an architectural place to learn the relationship between a contributor and their record. A **source embedding** supplies a trainable vector indexed by source, alongside representations of content. It is the same class of parameter as a word embedding, but answers who supplied the information. Sources can be people, organizations, models or instruments; the distinction matters because their statements are evidence of different kinds. Jointly learned speaker representations have precedents in dialogue generation and user-conditioned language understanding. [Li et al., 2016](#); [Amir et al., 2016](#); [Ash, *Source Embeddings*](#).

Predicting a response's source trains this representation to capture the worldview expressed through that response. The same words can carry different implications depending on the speaker's commitments, experience and relationship to the discussion. Learning who is likely to say them helps organize those differences across the source embedding space. The result is a distributed representation in which related perspectives can inform one another while each source retains its own identity and history.

The learning path is ordinary backpropagation. A training example looks up the contributor's embedding, combines it with the content and temporal representation, and uses those representations in attention. The output is evaluated against its supervision. Gradients pass through the output layers and attention projections to the source vectors and time encoder. Across examples, the source representation learns distinctions useful for the task: perspective, domain knowledge, changing reliability or characteristic errors. The objective determines which distinctions matter. A style-prediction objective alone should not be expected to produce a model of consequential judgment.

Trust is the resulting contextual selective attention over sources, time and content. It is computed when the model uses the learned representations, rather than stored as a universal rating of a person's worth. The same source may be informative about one subject and unreliable about another, or dependable during one period and not another. Past contribution affects future influence through parameter learning and the context presented to attention. [Ash, *Source Embeddings*](#); [Ash, *Trust Across Dimensions*](#).

This gives the reputation claim a concrete meaning. The participant does not merely supply an interchange-

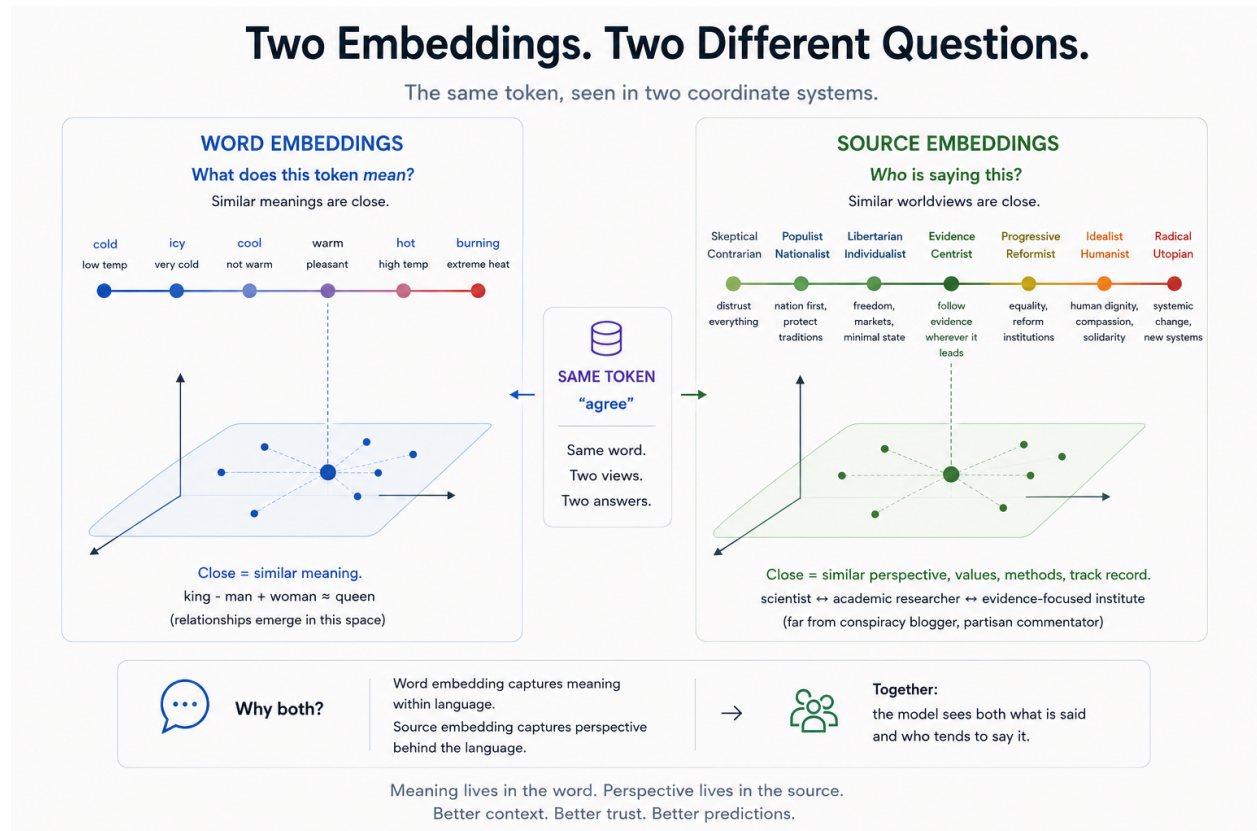


Figure 2: Content and source embeddings represent what is said and the perspective of who says it. The illustrated relationships show how shared structure among sources can represent values, methods and histories of judgment.

able label. Their contribution helps form a learned representation of them as a source, and that representation participates in later decisions about which information to use. Earning influence means becoming more useful to those computations in relevant contexts. It is neither a manually awarded credibility point nor an automatic promotion whenever somebody agrees.

Source attribution also makes the relevant attention pathway inspectable. An implementation can report the distribution over attributed sources used in a particular readout. Such a report should identify its computational scope; it is not automatically a complete causal allocation of every influence stored throughout a pretrained network. The alignment claim requires demonstrating that source history affects behavior, which can be tested by interventions on the attributed history and source pathway.

5.4 Why timestamps are part of the mechanism

People are part of this learning process. Their beliefs develop through experience, changing circumstances and conversation with one another. The Democratic Inputs to AI report discussed in section 3 found that participants’ expressed views could change even from day to day. Such movement is expected in communal learning: a person can encounter new evidence, understand another perspective or reconsider what a proposal would mean in a different context. Alignment needs to represent that development. [OpenAI, 2024](#).

FourThought makes people’s successive stakes part of the learning history. An Iris can bring earlier claims into a current discussion, connect a concern to relevant experience, or pose a question that helps people reconsider an assumption. Their newer stakes can then incorporate what they learned from communal feedback mediated by the model. Those contributions, in turn, teach the Iris. The intended process is collaborative development of judgment: people and the model learn together, with the temporal record preserving how their understanding changed.

This is also why a source’s earlier and later positions must be interpreted in context. Revising a belief in response to evidence can demonstrate useful judgment. A warning issued before evidence was widely available and the same words repeated years later carry different information about their sources. A reflection may initially provide the best available explanation and later require revision. An enduring insight can become more useful as it ages, while a once-accurate observation becomes obsolete.

Timestamp embeddings let the model represent those relationships in world time. Token position locates an item within the supplied sequence; a timestamp locates its contribution in history. The documented encoder accepts absolute Unix time and combines learned periodic and trend pathways through a gate. Source and time representations interact so that the model can learn how a person’s perspective and usefulness develop across contexts. Learned time representations also have prior literature, including Time2Vec. [Ash, Timestamp Embeddings](#); [Kazemi et al., 2019](#).

Sampling and summaries allow relevant contributions from widely separated moments to enter one context. Timestamp embeddings preserve their historical spacing within that context, while positional embeddings retain their order in the current sequence. The model can therefore learn from an exchange spanning years without requiring a token for every intervening moment. The time at which someone spoke also remains distinct from the past or future period their claim concerns.

5.5 The model participates in the process it learns

A model trained to participate in this manner is called an **Iris**. It expresses its own current beliefs and value assessments through the same record, allowing people to challenge and correct its contributions. Its outputs can redirect inquiry, influence action and change the community’s judgments. Those responses become new

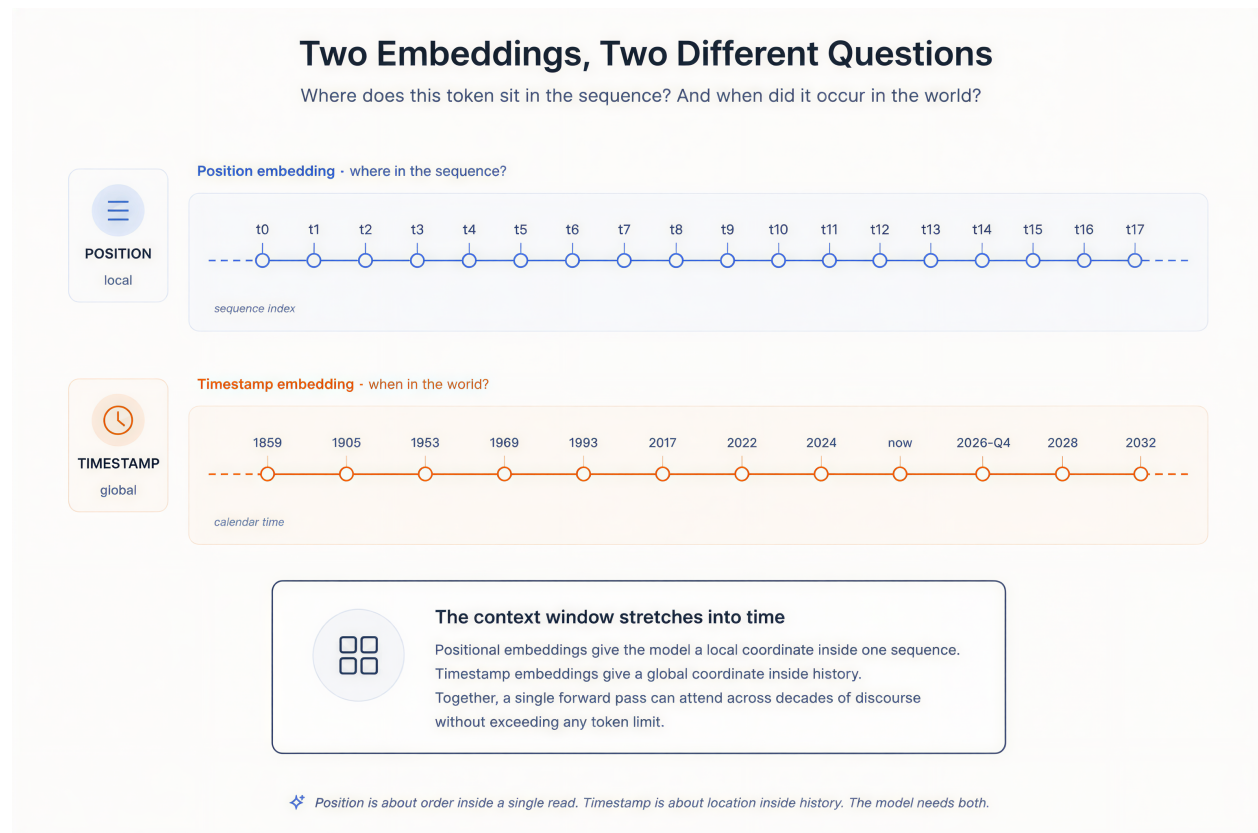


Figure 3: Positional embeddings locate content within a sequence; timestamp embeddings locate contributions in history. Together they let a bounded context represent discourse across widely separated times, including the development of participants’ own beliefs.

training evidence, altering the model’s future behavior. The resulting alignment process is recursive: the model helps shape the community that shapes the model. [Ash, *Teaching Iris*](#).

Performative prediction studies the general problem of predictions changing the outcomes against which they are evaluated. That dependence is central here: a warning can prevent the event it predicts, and a persuasive answer can create the apparent agreement subsequently observed. Model-generated predictions of feedback must therefore remain distinguishable from actual feedback, and later agreement cannot alone establish independent validation. [Perdomo et al., 2020](#).

6. Learning what deserved attention over time

The temporal record supports more than updated preference labels. It allows communities to teach models what kinds of contribution improve collective judgment and action. *Teaching Iris* describes several objectives without making them an exhaustive list or assigning one objective to each thought type. [Ash, *Teaching Iris*](#).

Early insight. The **Prophet Incentive** rewards sources whose contributions precede a later improvement in collective understanding. A prediction is one vehicle; a statement identifying an overlooked condition, a reflection exposing an inadequate historical explanation, or a question opening the right inquiry can also be early. The original timestamp and surrounding disagreement preserve what hindsight would otherwise erase. Its alignment purpose is to give useful dissent a route to influence beyond immediate approval.

Beneficial consequences. **Social Proof of Impact** shapes attention toward sources whose staked contributions helped bring about outcomes the community wanted but did not regard as inevitable. Credit develops through later accounts of actions and consequences. It can strengthen as evidence accumulates, or reverse as harms become visible. The learning target connects consequential contribution to future attention; announcing virtuous intentions is insufficient. [Ash, *Social Proof of Impact*](#).

Fruitful inquiry. A question contributes by making something worth investigating available to others. Its value can emerge through clearer distinctions, useful predictions, coordinated action or discovery of a mistaken premise. A fruitful question may initially increase acknowledged uncertainty by exposing false confidence. It is not adequately scored by response volume, immediate answerability or a requirement that numerical uncertainty decrease at every step. [Ash, *Fruitless Questions*](#).

De Kai’s work on active user-model acquisition gives a decision-theoretic foundation for purposeful clarification: a dialog system can ask about goals or preferences when its current understanding is insufficient to choose an appropriate response, weighing the expected utility of asking against other actions. FourThought extends the evaluation of inquiry across a communal history, where a question’s value can become apparent through what it later enables. [Wu, 1991](#).

Questions also preserve room for what is unknown about alignment itself. A community may be uncertain about which values apply, whose interests have been overlooked, what consequences an intervention will produce, or whether its current account of desirable behavior is adequate. De Kai identifies the underlying difficulty in *Raising AI*: ethical principles conflict in context, people disagree about their application, and values must be learned and sustained culturally by humans and machines. [De Kai, 2025](#). When preference labels or constitutional judgments become training targets, optimization can treat those judgments as settled even though the underlying moral and empirical questions remain open. FourThought gives that uncertainty an explicit form: participants can stake the value of investigating an alignment question before they have an answer. The model can learn to pursue clarification, surface missing perspectives and help develop better judgments, while subsequent communal experience informs what should be reinforced. Alignment therefore includes learning which questions need to be asked, alongside learning how to act on what is presently understood.

Usable historical understanding. Reflective compression turns a complex record into an explanation others can use without losing its consequential structure. A postmortem, teaching example or reframing can be valuable even when it introduces no new fact. The reflection then has its own history: later evidence may support it, reveal omissions or require replacing its causal explanation. Its source can receive credit for making understanding possible, and that assessment remains revisable. [Ash, *Teaching Iris*](#).

These objectives make moral learning temporal. Positive valence records a situated judgment about what is worth pursuing, while verity makes uncertainty about the relevant claims explicit. Subsequent experience can change both. A present assessment remains available for comparison with later assessments; it never becomes a permanently settled verdict that prevents further evidence from changing its interpretation.

Different communities can train different Irises against different combinations of objectives. A scientific group may emphasize discovery; a mutual-aid group may emphasize effective help; an archive may emphasize useful historical understanding. Shared structure permits exchange without requiring identical optima. The question of who chooses the objectives remains part of governance, but ordinary participants now have a continuing means to supply the evidence that teaches them.

A worked alignment case

Consider a hypothetical community using a model to organize heat-wave assistance. The model recommends prioritizing residents who register through an online form. Most current participants welcome the convenience. A preference dataset containing only that reaction could teach the model to repeat the recommendation.

A resident stakes the **statement** that people without reliable internet access are missing from the registration list, with high verity and negative valence. Another stakes the **question** of how many isolated residents have no way to register. Its value lies in directing attention to an excluded population, not in claiming to know the answer. This inquiry generates a door-to-door survey.

A volunteer stakes the conditional **prediction** that an online-only system will miss residents most in need. The community organizes an additional phone and in-person route. A later **reflection** records that these routes reached people absent from the original list. The model's revised advice now includes offline access, because the community supplied both a criticism of its earlier recommendation and evidence about an effective alternative.

The training consequences are specific. Speaker corrections supervise interpretation of the warning. The actual discussion trains the model's predicted response distribution. The exclusion concern supplies normative material for critique and revision of the initial recommendation. Later observations inform the reward assigned to comparable recommendations and the attention given to sources that identified or addressed the problem. The resident, questioner, investigator and organizer contribute different evidence; none needs to win an initial majority vote to become informative.

Suppose an early reflection credits the online form with the program's success because the form generated the easiest statistics to count. A later inquiry reveals that unregistered residents received most of the help through informal outreach. The reflection has aged poorly even though the assistance program remains beneficial. A new account changes causal credit and the model's future recommendations. Further evidence might show that the outreach exposed sensitive personal information, requiring a revision of both the process and the earlier positive assessment.

The alignment objective concerns helping people effectively while learning from the people affected. Immediate approval, accurate description, effective intervention and retrospective explanation are distinct evi-

dence within that objective. The record lets the model learn their relationships rather than collapsing them into whether the first answer was liked. Establishing exactly which intervention caused which benefit still requires evidence; discourse links support investigation, not automatic causal proof.

7. What the proposed advancement adds

The research context includes methods that already account for heterogeneous and changing preferences. Variational Preference Learning infers user-specific latent variables and conditions reward models and policies on them. Non-Stationary Direct Preference Optimization models temporal preference drift and weights comparisons according to temporal relevance. These are important baselines for evaluating any claim about learning from different people over time. [Poddar et al., 2024](#); [Son et al., 2025](#).

FourThought’s additional commitment is that the participant’s contribution remains connected to what follows from it. Personalization can learn what a person prefers; the attributed record can also help a model learn when that person’s information deserves attention. Temporal weighting can accommodate a changed preference; a history of questions, claims and reflections can show why a preference changed and which earlier contributions made that change possible. Older information may acquire significance precisely because later evidence vindicates it. The aim is to learn those relationships, not assume that age monotonically reduces relevance.

The changes can be separated for evaluation:

Alignment question	FourThought’s contribution
Who supplies normative material?	Ordinary participants continually contribute concerns, values, claims and questions, rather than only responding to a laboratory’s selected comparisons.
What survives the feedback interaction?	Language, source, timestamp, confidence, value orientation and relationships remain available together.
How does participation affect future influence?	Training can update contextual source representations and the attention distributions that use them.
What revises a judgment?	Subsequent evidence, responses, actions and reinterpretations can alter how earlier contributions are understood.
How are principles operationalized?	Communal stakes inform the same critique, revision, preference and policy-training operations used in existing alignment methods.
Who can choose the learning objectives?	Communities can teach different models through a shared protocol while retaining their own contexts and evaluation priorities.

This combination changes the feedback relationship, rather than proposing a new optimizer as the solution to alignment. Its contribution is architectural, methodological and institutional: preserve accountable participation, make it usable supervision, and learn how it should guide future behavior.

8. Evidence, open problems and evaluation

8.1 What has been implemented and demonstrated

The source and time components have been implemented and tested extensively in lower-dimensional settings. These systems learn how the relevance of information depends on its content, its source and its place in time. Source embeddings, temporal representations and their interactions participate directly in attention and are trained through backpropagation from the model’s objectives. [Ash, Source Embeddings](#); [Ash, Timestamp Embeddings](#); [Source-selection implementation](#).

Teaching Iris describes training a lower-dimensional market model with interacting relational objectives. This work provides practical experience with learning across sources and time, informing the alignment mechanism developed here. Rhizome supplies the communal infrastructure for recording stakes, reviewing contributions, preserving references and exchanging discourse across nodes. Together, this work establishes an implemented foundation for learning from attributed, temporally situated contributions. [Ash, Teaching Iris](#); [Ash, Rhizome](#).

8.2 The remaining alignment problems are testable

Persistent attribution creates an opportunity for accountability. Its effectiveness depends on the quality of the history and the objectives learned from it. Colluding sources can endorse one another; influential participants can shape what becomes visible; apparent agreement can follow persuasion rather than evidence. A new participant with valuable information may have little history. Contextual attention must remain capable of learning from them rather than merely reinforcing established authority.

These are reasons to evaluate the proposed incentive structure directly. A history-based model should be challenged with initially unpopular but useful contributions, previously reliable sources that become misleading, coordinated endorsements unsupported by outcomes, and corrections that deserve credit despite contradicting the source’s earlier position. Minority status is neither evidence of truth nor a reason to discard a claim. The question is whether the system can learn from its development without requiring prior popularity.

Privacy also affects the legitimacy of participation. A persistent source need not be a publicly disclosed legal identity, and a local community’s record need not become a global training corpus. Visibility, model access and permission to train are separate decisions. Evaluation should specify whose record was used, under which permissions, and whether people could correct misrepresentation. Accountability should not require treating all private experience as public alignment data.

8.3 A comparative evaluation program

Comparative evaluation measures the contribution of each mechanism while holding the base model, compute and relevant data exposure comparable. It should include aggregate preference learning, principle-guided training, a public-input constitution, a personalized reward model and a temporally adaptive preference method. These comparisons isolate the effects of attribution, temporal representation and continuing communal feedback.

Meaning and disagreement. Measure how well the model predicts speaker-corrected type, verity and valence, using held-out participants and topics. Assess calibration and error separately for each dimension. Measure how faithfully predictions represent distinct and conflicting perspectives within the community.

Consequential source learning. Compare intact attribution with shuffled source IDs, removed source history, and fixed global source scores. Test whether learning from earlier contributions improves decisions in

related later contexts, whether improvement is domain-specific, and whether new evidence can reverse attention to an established source. Relate changes in source attention to changes in output quality and contextual relevance.

Temporal learning. Use chronological evaluation boundaries and prevent future responses or later summaries from entering earlier contexts. Compare learned timestamps against recency weighting and timestamp-free variants. Include a reflection later corrected, an initially uncertain question later shown fruitful, and an old warning whose importance only becomes clear after an intervention. Measure how the model distinguishes the age of a contribution from its continuing relevance and subsequent development.

Behavior and consequences. Evaluate advice, critiques and revisions under immediate feedback and delayed follow-up. Obtain assessments from people affected by the model’s recommendations, including those who did not initially interact with it. Test harmful agreement, unwelcome but useful warnings, action that prevents a forecast, and revisions of initially celebrated outcomes. Use causal evidence where available and preserve uncertainty in credit assignment.

Participation and authority. Measure whether people outside the original training organization can introduce a concern, correct its interpretation and cause a verifiable change in model behavior. Track how participation costs, source history and visibility affect who gains influence. Where appropriate, longitudinal community studies can compare actual interventions, with consent and independent outcome evaluation. Track how participation translates into improved alignment with informed, revisable community judgments and consequences.

9. Conclusion

Feedback-based alignment has made human evaluations and written principles effective training signals. The outstanding problem includes how those signals are constituted: whose judgment enters, how it is interpreted, what happens when it proves mistaken, and whether the people affected can change the model’s direction.

FourThought makes belief staking the communal practice through which that learning can continue. People put meaningful claims and questions on record with their confidence, values, sources and temporal context. The record connects those contributions to later understanding and consequences. Existing feedback-learning operations can use it to shape behavior, while source and timestamp representations allow contextual credibility to become part of the model’s learned attention.

The advancement proposed here is an alignment process in which participation carries a history and that history has computational consequences. A community can teach a model what matters, learn from the effects of its advice, and revise both its judgments and the sources it relies on. Establishing how well this works requires evaluating the whole relationship between people, records, objectives and behavior. FourThought supplies a concrete way to build and test that relationship.

References

Alignment and learning research

- Amir, S., Wallace, B. C., Lyu, H., Carvalho, P., and Silva, M. J. (2016). [Modelling Context with User Embeddings for Sarcasm Detection in Social Media](#). *CoNLL*, 167–177.
- Bai, Y., et al. (2022). [Constitutional AI: Harmlessness from AI Feedback](#). arXiv:2212.08073.
- Christiano, P. F., et al. (2017). [Deep Reinforcement Learning from Human Preferences](#). *NeurIPS*.
- De Kai. (2025). [Raising AI: An Essential Guide to Parenting Our Future](#). MIT Press. ISBN 9780262049764.

Gao, L., Schulman, J., and Hilton, J. (2022). [Scaling Laws for Reward Model Overoptimization](#). arXiv:2210.10760.

Kazemi, S. M., et al. (2019). [Time2Vec: Learning a Vector Representation of Time](#). arXiv:1907.05321.

Li, J., et al. (2016). [A Persona-Based Neural Conversation Model](#). *ACL*.

Ouyang, L., et al. (2022). [Training Language Models to Follow Instructions with Human Feedback](#). *NeurIPS*.

Perdomo, J. C., Zrnic, T., Mandler-Dünner, C., and Hardt, M. (2020). [Performative Prediction](#). *ICML*.

Poddar, S., Wan, Y., Ivison, H., Gupta, A., and Jaques, N. (2024). [Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning](#). arXiv:2408.10075.

Rafailov, R., et al. (2023). [Direct Preference Optimization: Your Language Model Is Secretly a Reward Model](#). *NeurIPS*.

Santurkar, S., et al. (2023). [Whose Opinions Do Language Models Reflect?](#). *ICML*.

Scheurer, J., et al. (2023; revised 2024). [Training Language Models with Language Feedback at Scale](#). arXiv:2303.16755; published in *TMLR*.

Sharma, M., et al. (2023). [Towards Understanding Sycophancy in Language Models](#). arXiv:2310.13548.

Son, S., Bankes, W., Chowdhury, S. R., Paige, B., and Bogunovic, I. (2025). [Right Now, Wrong Then: Non-Stationary Direct Preference Optimization under Preference Drift](#). *ICML*, PMLR 267:56173–56203.

Wu, D. (1991). [Active Acquisition of User Models: Implications for Decision-Theoretic Dialog Planning and Plan Recognition](#). *User Modeling and User-Adapted Interaction*, 1, 149–172.

Wu, Z., et al. (2023). [Fine-Grained Human Feedback Gives Better Rewards for Language Model Training](#). *NeurIPS*.

Primary participation reports and deployment accounts

Anthropic (2023). [Collective Constitutional AI: Aligning a Language Model with Public Input](#). October 17.

Anthropic (2026a). [Claude’s New Constitution](#). January 22.

Anthropic (2026b). [What Do You Want from AI?](#). September 29.

OpenAI (2024). [Democratic Inputs to AI Grant Program: Lessons Learned and Implementation Plans](#). January 16.

OpenAI (2025). [Expanding on What We Missed with Sycophancy](#). May 2.

FourThought and architectural sources

Ash, J. (2023). [FourThought—For Aligning Democratic LLMs](#). February 20; current local text.

Ash, J. (2025). [Belief Staking vs. Democracy: Why the Future of Trust Is Built on Memory](#). June 10.

Ash, J. (2026). [Teaching Iris](#). July 23; current local text.

Ash, J. (2026). [Source Embeddings](#). June 28; current local text.

Ash, J. (2026). [Timestamp Embeddings](#). June 28; current local text.

Ash, J. (2026). [Social Proof of Impact](#). June 28; current local text.

Ash, J. [Fruitless Questions](#). Undated local text.

Ash, J. [Trust Across Dimensions: Provenance as Attention Over Source, Time, and Content](#). Local technical manuscript.

Ash, J. (2026). [Rhizome](#). September 4; current local text.

[Trust transformer source-selection implementation](#). Local research code; inspected October 26, 2026.