Bonsai: Interpretable Tree-Adaptive Grounded Reasoning

Kate Sanders, Benjamin Van Durme

Johns Hopkins University
{ksande25, vandurme} @jhu.edu

Abstract

To develop general-purpose collaborative agents, humans need reliable AI systems that can (1) adapt to new domains and (2) transparently reason with uncertainty to allow for verification and correction. Black-box models demonstrate powerful data processing abilities but do not satisfy these criteria due to their opaqueness, domain specificity, and lack of uncertainty awareness. We introduce BONSAI, a compositional and probabilistic reasoning system that generates adaptable inference trees by retrieving relevant grounding evidence and using it to compute likelihoods of sub-claims derived from broader natural language inferences. BONSAI's reasoning power is tunable at test-time via evidence scaling and it demonstrates reliable handling of varied domains including transcripts, photographs, videos, audio, and databases. Question-answering and human alignment experiments demonstrate that BONSAI matches the performance of domain-specific black-box methods while generating interpretable, grounded, and uncertainty-aware reasoning traces.

Introduction

Human professionals often write full documents describing the veracity and scope of individual claims, while many AI systems consider them to simply be true or false. To be useful in practical settings, reasoning systems must be able to model concerns like subjectivity, epistemic uncertainty, and ambiguity in natural language statements, and should be able to identify which portions of the statements these concerns apply to. Furthermore, these systems should be robust to knowledge sources of different modalities, as in many settings a grounding source may be a research report, a photograph, or a news article with embedded video clips. With these ideas in mind we introduce BONSAI, an adaptable reasoning tree generator for transparent, grounded, and probabilistic inference. BONSAI introduces a set of key design choices that enable sophisticated interpretable reasoning.

First, BONSAI extends the "evidence extraction" paradigm – in which natural language summaries of complex or out-of-distribution source documents are generated as data to reason over (Li et al. 2024) – to multimodal content. It accomplishes this by mapping non-textual data to evidence banks of natural language observations which it draws from during reasoning. BONSAI applies contextual conditioning to generated observations to mitigate the well-documented issue of perspective ambiguity in multimodal domains (Zur et al. 2024).

BONSAI accounts for uncertainty and subjectivity in data by replacing categorical labels used in traditional claim verification ("true", "false", etc.) with scalar likelihood scores. We introduce an iterative approach to scalar likelihood score calculation using retrieved evidence samples as explanatory conditional variables, inspired by Tversky and Kahneman (1974)'s "anchoring and adjustment" framework for human judgments. This approach integrates naturally with chain-of-thought (Wei et al. 2022) and allows BONSAI to behave as an *evidence-grounded* adaptable prediction system (Mohri and Hashimoto 2024; Jiang, Liu, and Van Durme 2025) that can restructure its output depending on risk threshold.

Finally, BONSAI outputs grounded explanations for each sub-claim judgment, which may be propagated upwards to generate likelihood scores and explanations for any set of sub-claims. Since BONSAI decomposes claims into tree structures (illustrated in Fig. 1) this enables straightforward human analysis and quick correction of intermediate sub-claims, in contrast to individually assessing every atomic claim, one-by-one. Using tree structures, contextualized evidence extraction may be computed at arbitrary levels of granularity. Through this we adapt test-time search scaling (Zhao, Awasthi, and Gollapudi 2025) to facilitate flexibility in performance-compute tradeoffs with test-time *evidence* search scaling.

In summary, BONSAI is a transparent and probabilistic multimodal reasoning system that promotes three key ideas: (1) Mapping raw data to contextualized natural language observations enables high-performance modality-agnostic reasoning, (2) assigning sub-claims with probabilistic scalars enables adaptable prediction and nuanced reasoning over ambiguities, and (3) coupled with (1) and (2), a tree-based decomposition structure can enable impactful compute scaling and easy human-in-the-loop interpretation and corrections. Through experiments, we demonstrate that BONSAI enables state-of-the-art performance on both single- and multimodal tasks such as EntailmentBank (Dalvi et al. 2021) and TVQA (Lei et al. 2018), while crucially providing a fully grounded, human interpretable thought process.

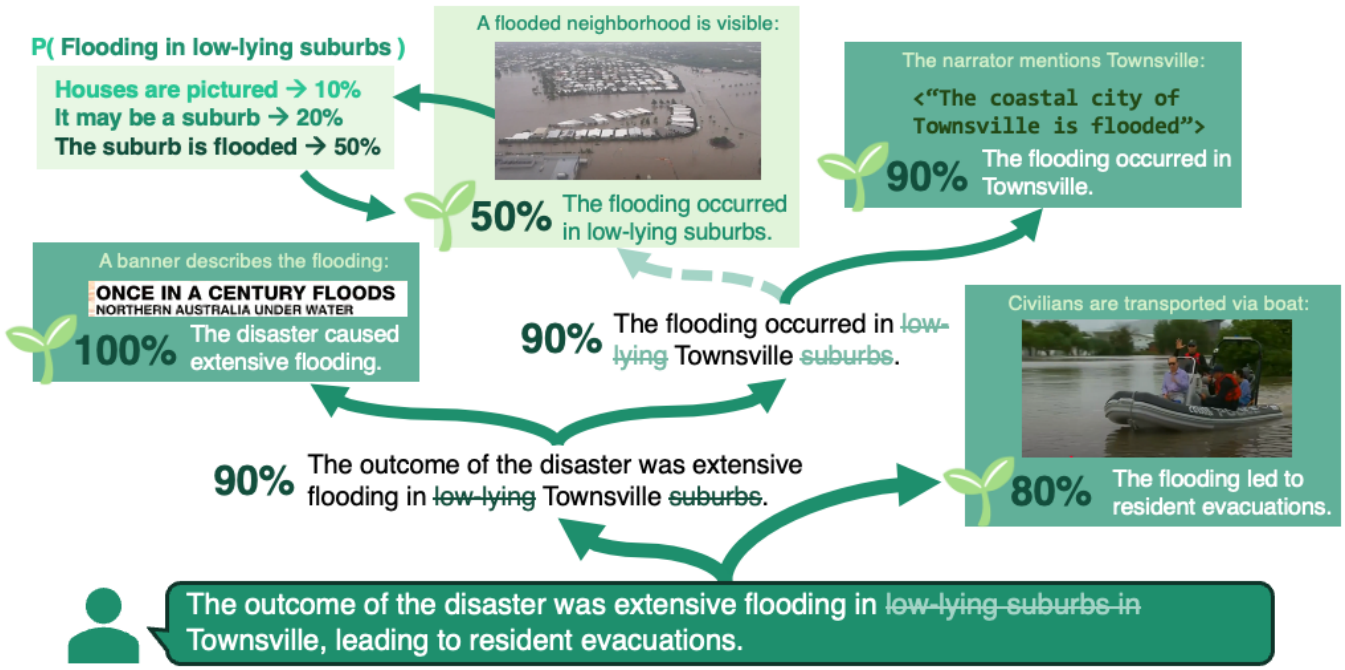

Figure 1: A reasoning tree over a news video as a grounding source. BONSAI recursively decomposes natural language statements about data into small, verifiable pieces. It uses retrieved evidence samples from multimodal knowledge sources to iteratively score these pieces in terms of how likely each piece is. This procedure results in grounded likelihood scores for leaves of compositional tree structures representing the original claim, alongside natural language explanations. Low-scoring branches of a reasoning tree may then be pruned, shown by the strikethrough text in the original statement and sub-claims.

Related Work

Probabilistic Reasoning with LLMs

Earlier research in probability estimation has explored fine-tuning models on human probability judgments (Chen et al. 2019). Recent work has investigated if LLMs can perform probabilistic calculations, such as estimating percentiles and basic probability computations (Paruchuri et al. 2024), and even update belief states given new information (Qiu et al. 2024). The BIRD framework leverages LLM judgments for downstream probability inference (Feng et al. 2024), and Piriyakulkij et al. (2024) and Mo and Xin (2024) leverage Monte-Carlo algorithms. Notably, some approaches consider graph-centric probabilistic reasoning, either using explicit knowledge graphs (Li, Zhang, and Liu 2024) or graphs extracted via chain-of-thought reasoning (Razghandi, Hosseini, and Baghshah 2025). A growing body of work considers LM uncertainty quantification (Xiao et al. 2022; Stengel-Eskin, Hase, and Bansal 2024), some directly via prompting (Tian et al. 2023; Xiong et al. 2023). Some methods incorporate conformal prediction by considering object sets in their probabilistic reasoning (Ozturkler et al. 2022). Conformal prediction produces answer sets or intervals with assigned correctness probability scores (Angelopoulos, Bates et al. 2023). Many applications to language modeling have been identified in recent years: Quach et al. (2023) propose a framework for conformal language modeling, Mohri and Hashimoto (2024) leverage conformal prediction for correctness guarantees, and Jiang, Liu, and Van Durme (2025) introduce a pragmatics-inspired framework for factuality-specificity tradeoffs.

Transparent Reasoning

While chain-of-thought (Wei et al. 2022) and related approaches are critical to understanding LLM explanations, alongside the complementary vein of research in reasoning model traces (Jaech et al. 2024), these methods generally lack the trustworthiness of fully transparent reasoning approaches (Lanham et al. 2023; Yeo et al. 2024). Such lines of work include entailment tree generation, in which claims are recursively decomposed and verified through entailment using an underlying knowledge source (Weir, Clark, and Van Durme 2022), Proof of Thought, a first-order logic-inspired LLM approach (Ganguly et al. 2024), other tree- or graph-based reasoning methods (Luo et al. 2023; Mei et al. 2024) such as those leveraging Monte-Carlo Tree Search (Gao et al. 2024b), and other approaches that draw more direct inspiration from chain-of-thought (Lyu et al. 2023; Chen et al. 2024b).

Claim Verification

LLMs have enabled significant progress in the field of fact-checking (Bekoulis, Papagiannopoulou, and Deligiannis 2021; Dmonte et al. 2024) on a variety of benchmarks including SciFact (Wadden et al. 2020), FEVER (Thorne et al. 2018), FactScore (Min et al. 2023), and X-FACT (Gupta and Srikumar 2021). Many approaches (Chen et al. 2024a) center on claim decomposition (Wanner et al. 2024), and

retrieval-augmented generation (Gao et al. 2024a) is often applied in claim verification settings, sometimes with notable success (Xu et al. 2024; Kao and Yen 2024). Recent work has addressed evidence extraction as an intermediate step in claim verification (Cao et al. 2024; Li et al. 2024). Other work focuses on accurate attributions of claims generated by the models themselves (Press et al. 2024; Weller et al. 2024). Srikanth and Rudinger (2025) apply an iterative inference procedure over decomposed evidence to traditional and defeasible NLI. ClaimVer (Dammu et al. 2024) grounds decomposed claims to source documents with generated explanations. Our work additionally incorporates multimodality and human-interpretable, sub-claim-level probability scores and explanations.

BONSAI Reasoning Tree Generation

A BONSAI reasoning tree begins with a single natural language statement that serves as the "root", which is recursively decomposed into (usually binary) sub-claims until the claims reach an atomic state. These decompositions are included in the trace and serve as the branches of the tree. A leaf is made up of an atomic sub-claim paired with (1) the top-k most relevant evidence pieces from the grounding data (documents, videos, databases, etc.), (2) a sub-claim likelihood score, and (3) a natural language explanation detailing how the evidence was used to compute that score. These scores and explanations may be propagated up the tree branches, which we explore in Section . Below, we provide further details regarding the remaining aspects of tree construction.

Claim Decomposition

BONSAI produces a hierarchical representation of individual claims, similar to the structure of an incomplete entailment tree. We begin with the decomposition of the initial hypothesis into compositionally entailing premises. We use GPT-4o (Hurst et al. 2024) to compute decompositions and provide one example decomposition for guidance, specifying the syntactic nature of the decompositions. This repeats until individual sub-claims reach an atomic state (in this setting, no longer syntactically decomposable) or the depth of the tree reaches limit k. Although it is not encouraged, occasionally GPT outputs more than two sub-hypotheses for a given decomposition. We allow and account for this behavior as it generally occurs in scenarios where > 2 premises is appropriate, e.g., "The hurricane affected Barbuda, Cuba, and Haiti" $\rightarrow$ "The hurricane affected Barbuda" + "The hurricane affected Cuba" + "The hurricane affected Haiti".

Evidence Extraction and Retrieval

BONSAI performs reasoning over prespecified single- or multimodal grounding sources. To enable robust modality- and domain-agnostic reasoning, BONSAI constructs an evidence bank of natural language observations derived from the grounding source instead of using domain-specific representations. These observations are individually mapped to specific spans of the grounding source, enabling BONSAI reasoning to be fully grounded. These spans are determined offline and uniformly: Text is split into partially-overlapping windows of 6 to 12 lines depending on length. Images are passed in individually with no preprocessing. Between 1 and 10 frames are sampled from videos, depending on length. ASR is performed on audio content, and the extracted text is partitioned as a regular text document. We prompt LLMs and MLLMs to extract these observations, asking for a set of captions over small samples of the source data to ground observations in specific portions of the grounding source. These prompts may include context to improve reasoning ability (discussed further in Section).

Retrieval over this generated evidence bank is necessary, as it becomes intractable to consider all evidence factors in most scenarios when computing the likelihood of a sub-claim. So, we use a heuristic to identify the most promising evidence factors. We pass a set of pre-extracted natural language factors into a cross-encoder model trained on the MS-MARCO dataset (Bajaj et al. 2016) to identify the top-k evidence pieces. Increasing k generally improves model reasoning at the cost of compute. $k = 3$ to $k = 10$ are often practical.

When working with temporal data, this evidence extraction approach notably eliminates temporal (or other ordering) information, which may be critical in applications like video understanding benchmarks. To account for this, we explore the application of an adjustment in which we prepend approximate temporal metadata to retrieved evidence snippets, order the evidence presented to the probability scorer temporally, and include a line to the probability scorer indicating that temporal information may be present.

Sub-Claim Scoring

Given some sub-claim and a set of observational evidence factors $\{f_e^{(0)}, ..., f_e^{(n-1)}\}$, we leverage LLM agents as "knowledge experts" from which we elicit probabilistic judgments about the sub-claim. Following research in economics (Tversky and Kahneman 1974), we frame this problem as an "anchoring and adjustment" task through which we leverage LLMs' demonstrably strong chain-of-thought reasoning.

We provide an LLM with an initial guidance rubric in which we carefully describe the probabilistic scale from 0 to 100 in human-interpretable terms. We provide the agent with an artificially constructed evidence factor, or initial premise, and elicit an initial "anchoring" probability score following the rubric. The artificial factor or premise may be constructed by generating a brief summary of the input observations. This produces an initial prior anchor score that is conditioned on a generic scenario within the scope of the specific world state. From here, the model is presented with the real evidence factors and is instructed to iteratively adjust its initial anchor score, factor-by-factor. Sample observations are shown in Figure 2, compared against human assessments of the video.

Inference with BONSAI

Given BONSAI's detailed reasoning traces, there are multiple ways to perform inference depending on the task and compute. In this section, we consider a probabilistically sound approach to generate individual likelihoods as well as a counterfactual reasoning method for multiple choice question-answering. We also touch on a widely applicable test-time search scaling method for improved performance.

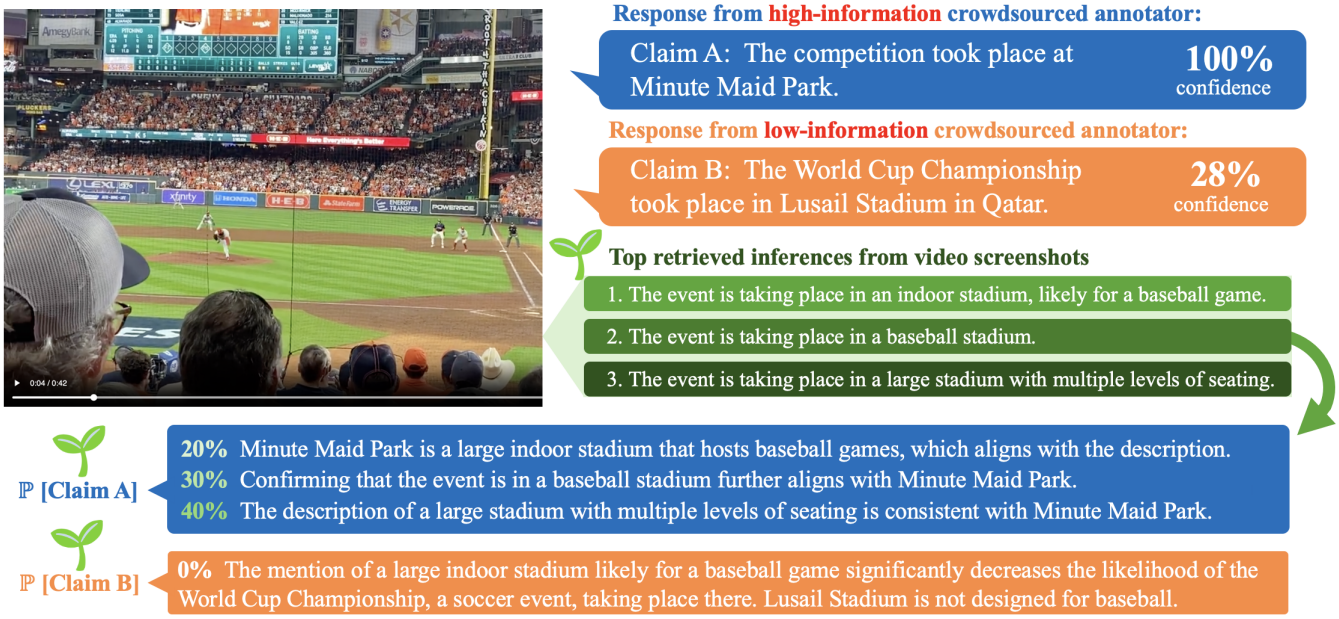

Figure 2: For most tasks, human performance varies depending on background knowledge. Two humans were given a set of candidate descriptions for the video on the left, and pictured on the right are their answers and confidence scores. Illustrated below the responses, BONSAI retrieves its top three generated video observations and uses them to score these claims, in the positive case by iteratively updating their likelihood scores and providing explanations.

Complete Probabilistic Inference

We may leverage the entailment tree structure of BONSAI's reasoning traces to compute the conditional relationships between sub-claims. Let our observational evidence for sub-claim e be $O_e = \{f_e^{(0)}, f_e^{(1)}, ..., f_e^{(n-1)}\}$. If factors A and B syntactically compose H, then we make the assumption $P(H) = P(A \cap B)$, and consequently, $P(H|O) = P(A|B,O)\ P(B|O) = P(B|A,O)\ P(A|O)$. As evidence factors and sub-claims are both natural language strings, we may simply view the computation of $P(A|B,O)$ as $P(A|\{f_A^{(0)}, f_A^{(1)}, ..., f_A^{(n-1)}, B\})$, with no modifications to the likelihood computation method.

Letting A entail (C, D), then this propagation may operate recursively as $P(H|O) = P(C|D,B,O)P(D|B,O)P(B|O)$. The remaining issue is how to select which variable to condition on (A vs. B, or C vs. D) in such a decomposition. Ideally, both would be computed and expert aggregation would be performed, but this approaches exponential complexity as the structure of the decompositional tree grows. Therefore, we only compute either $P(A|B)$ or $P(B|A)$. This full process is detailed in Algorithm 1.

Counter-Factual Reasoning

Results show that sampling multiple answers and reasoning over them can result in higher performance on tasks than direct inference with complex reasoning models (Zhao, Awasthi, and Gollapudi 2025). Similarly, many domains directly facilitate the comparison of multiple options, such as multiple choice QA. In such scenarios where multiple options are being considered, we introduce two primary additions

Algorithm 1: Probabilistic inference, INFER

Require: Decomposition tree root r, evidence factor set $\mathcal{F}$ (the set of all observational evidence pieces from the grounding source), additional conditional factors $\mathcal{F}_c$ (from other propagated branches), and anchor factor f_a.
Ensure: Likelihood of root factor r conditioned on evidence, $P(r|\mathcal{F}, \mathcal{F}_c, f_a)$.

if CHILDREN $(r) = \emptyset$ **then** ▷ Check if the current root is a leaf of the tree.
  $\mathcal{F}' \leftarrow \mathcal{F}_c \cup$ RETRIEVE$_\varepsilon$ $(r, \mathcal{F})$ ▷ Combine evidence and propagated factors.
  $P_0 \leftarrow$ ANCHOR (f_t, f_a) ▷ Compute base probability score.
  $\mathcal{F}^+ \leftarrow \emptyset$
  for $f' \in \mathcal{F}'$ **do**
    $\mathcal{F}^+ \leftarrow \mathcal{F}^+ \cup \{f'\}$
    $P_0 \leftarrow$ ADJUST $(P_0, r, \mathcal{F}^+)$ ▷ Update probability for each piece of evidence.
  end for
else
  $A, B \leftarrow$ CHILDREN (r)
  $P_A \leftarrow$ INFER $(A, \mathcal{F}', \mathcal{F}_c \cup B, f_a)$ ▷ Recurse on child branches, propagate factors.
  $P_B \leftarrow$ INFER $(B, \mathcal{F}', \mathcal{F}_c, f_a)$
  $P_0 \leftarrow P_A \cdot P_B$ ▷ Compute root probability.
end if
return P_0

to BONSAI to enable strong counterfactual reasoning. (1) In many cases, these different sampled answers may share similar components of their respective claim decomposition trees. We prune leaf nodes of answer trees that are inherently entailed by the other hypotheses using a cross encoder trained on SNLI (Bowman et al. 2015), so that when making a decision, we consider the primary factors that distinguish the options from one another. (2) In cases where we know one of the inferences are true (for example, multiple choice), we can provide this information as conditional context to limit the world space being considered by the probability scoring system: Instead of measuring the general probability of a hypothesis, we can directly compute the relative likelihoods of different options conditioned on the fact that one is true.

Given a set of counterfactual reasoning trees, there are multiple methods to select a final answer. While most probabilistically sound, we find that constructing final probability scores using the process detailed in Section (or similar approaches) penalizes more complex answers with more leaves. Therefore, in practical comparison settings we opt to take the average score across all leaves. However, this suffers from not modeling any interdependencies between leaves, as well as not sufficiently penalizing scores that accrue one or more low-probability leaves. To remedy these issues (at the cost of transparency) we also consider an LLM-based "judge" method that outputs a final answer based on the collection of leaf sub-claims and their likelihood computed scores.

Test-Time Evidence Search Scaling

Depending on the complexity of the underlying data, the generic evidence extracted through may not be sufficient to output a high probability score for any complex claims about the content. In such a case, this will be demonstrated by the consistently low resulting probability scores in a multiple-choice or multi-inference setting. Given these scores, the system may choose to engage in a second (or nth) round of evidence extraction and claim re-scoring. In this re-extraction, claims or sub-claims constructed during the claim decomposition step () may be passed in as contextualizing information to the extractor, resulting in more specific and topical observations. While increasingly computationally expensive, as the context used for evidence extraction grows more specific the confidence of the model correspondingly improves. We illustrate the benefit of such an approach in Section .

Experiments

We evaluate BONSAI on four tasks. We first consider the quality of the proposed calibration approach, using a dataset of human-scored ambiguous images (Section). Then, we test the full reasoning system on traditional single- and multi-modal question-answering tasks (Section). We finally evaluate the system on a multimodal inference task that involves reasoning over ambiguity (Section).

Likelihood Calibration

We first characterize the quality of probability judgments produced by BONSAI by comparing them against human probability judgments in a low-information setting.

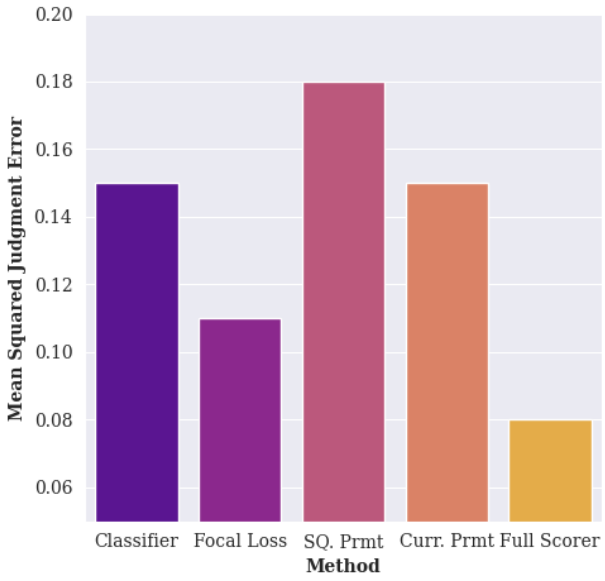

Figure 3: Agreement of different uncertainty quantification approaches compared to human likelihood judgments on a visual classification task. Unlike basic probability scoring prompts (SQ. Prmt and Curr. Prmt), BONSAI's "anchor and adjust" evidence-focused approach (Full Scorer) outperforms traditional uncertainty quantification methods with a fine-tuned classifier (Focal Loss).

Task We evaluate BONSAI on SQUID-E (Sanders et al. 2022), a collection of ambiguous images sampled from videos of everyday events like weddings and medical procedures. Alongside ground truth labels of these events, the images are labeled with human probability judgments on a 0-100 scale that quantify how likely people believe the images depict these ground-truth events. We evaluate on the full set of 3,600 human-labeled images. We evaluate the MSE of BONSAI's probability judgments and the median human judgments for each image in the test set. We compare BONSAI against a traditional visual event classification network, this network combined with the most effective model calibration technique identified in the SQUID-E publication, focal loss (Lin et al. 2017), alongside alternative LLM-driven probability solicitation approaches.

Setup We use Molmo 7B (Deitke et al. 2024) for image evidence extraction, and GPT-4o for probability scoring. As the hypotheses in this task are already atomic (essentially the one- or two-word description of the event) we do not further decompose the hypothesis. We test two prompts alongside the traditional classification approaches (Classifier, Focal Loss) and BONSAI's probability scoring algorithm (Full Scorer): SQ. Prmt is a prompt-adapted version of the original instructions given to human annotators for SQUID-E, and Curr. Prmt is a prompt-adapted version of the probability scoring outlines used in BONSAI, but without the anchoring-and-adjustment portion used with extracted visual evidence.

Model	Transp.	EB/ARC Acc.	Model	Transp.	TVQA Acc.
NELLIE	✓	71.4	VideoChat2	✗	40.6
TreeWise	✓	79.2	TV-TREES	✓	49.4
GPT-3.5*	✗	88.7	MiniGPT4-Video	✗	54.2
GPT-4o*	✗	96.0	IG-VLM	✗	57.8
BONSAI	✓	87.7	**BONSAI**	✓	65.5
BONSAI$_{J\text{-}4o}$	✓	**95.6**	**BONSAI**$_{J\text{-}4o}$	✓	**68.8**

Table 1: BONSAI performance on traditional multiple-choice tasks, EntailmentBank and TVQA, compared against other transparent and black-box methods (all zero-shot). GPT-3.5 and GPT-4o are run on standard ARC(*), not EntailmentBank with the WorldTree grounding source. On both tasks, BONSAI with basic scoring already outperforms other transparent approaches, and using a 4o judge (BONSAI$_{J\text{-}4o}$) it achieves another performance boost.

Model	Transp.	MV Acc.
LLaVA-Next-7B	✗	61
InternVL 2.5-8B	✗	83
Qwen VL 2-7B	✗	87
BONSAI	✓	76
BONSAI$_{SL}$	✓	**89**
Human (Avg.)	✓	93
Human (Best)	✓	100

Table 2: Results of the ambiguous video understanding task against SoTA video models. BONSAI's results demonstrate the power of evidence scaling, matching the performance of SoTA black-box video models.

Results Agreement against ground-truth human scores is shown in Fig. 3. While the likelihood rubric used in BONSAI on its own outperforms a basic likelihood scoring prompt and matches the alignment of fine-tuned classifier logits, it underperforms compared to the uncertainty quantification approach. Only BONSAI's anchor-and-adjust method aligns with humans better than all other methods, indicating the efficacy of a more in-depth probability scoring method when considering human alignment.

Traditional Multiple Choice QA

We explore whether BONSAI maintains high-quality performance on traditional question-answering tasks in text and multimodal domains.

Tasks Following the experimental setup of existing reasoning trace generation work, we evaluate on EntailmentBank (Dalvi et al. 2021) and TVQA (Lei et al. 2018). EntailmentBank consists of multiple choice questions taken from middle school science curricula (from the ARC dataset (Clark et al. 2018)), and provides an evidence bank of factual science statements from the WorldTree dataset (Jansen et al. 2018) to use as supporting evidence for answers. TVQA consists of multiple choice questions about popular TV shows, and provides episode clips and dialogue transcripts as supporting evidence to draw from. We sample 1,000 questions from each dataset (requiring sampling from the training set for EntailmentBank). For both datasets we use mean accuracy as the evaluation metric, and we compare BONSAI against contemporary end-to-end models designed for the tasks as well as architecturally similar neuro-symbolic methods. For EntailmentBank, there is not a naturally fair black-box comparison to the transparent systems, as the WorldTree corpus includes 11,941 natural language facts, or over 150K tokens. To approximate, we run GPT on the standard ARC benchmark, in which they use their parametric knowledge instead of a grounding source.

Setup For both datasets, we use GPT-4o to convert each answer choice into a natural language hypothesis. For example, the question answer pair "Q: At what temperature does ice melt? A: 72 degrees Celsius" would become "Ice melts at a temperature of 72 degrees Celsius". For EntailmentBank, we use the provided WorldTree factbase as our precomputed evidence collection. For transcript text inference on TVQA, we use GPT-4o with context windows of 6 lines of dialogue, and for vision we use Molmo 7B with approx. 10 sampled frames per video. For conditioning context for probability scoring in EntailmentBank we use the original question with the framing "someone is asking the question, ...", and for TVQA we extract a short transcript summary with GPT-4o. For both tasks we use the counterfactual reasoning approach detailed in Sec. . We use the two scoring alternatives in Section : Aggregation of leaves (via multiplication for ARC and mean for TVQA, as ARC has similar leaf counts per answer while TVQA does not), and a GPT-4o judge that takes leaves and outputs the likeliest answer (written BONSAI$_{J\text{-}4o}$).

Results Results are shown in Table 1. On TVQA, the raw BONSAI scorer outperforms other methods developed for the benchmark, indicating a strong ability to reason over multiple modalities simultaneously while still grounding its output in specific content from its input data. On EntailmentBank, the raw BONSAI scorer performs competitively, which illustrates that while BONSAI can generalize beyond text-centric tasks, it can still retain appropriate performance on more traditional text benchmarks. Using the 4o scoring judge it is able to roughly match the performance of black-box GPT-4o while being grounded in a set of specific scientific facts, instead of being able to freely rely on parametric knowledge.

The LLM-assisted probability aggregation is effective on both tasks. This is reasonable, as the alternate methods of

Claim	Evidence	Scoring Output
The first photons will be captured on Webb's primary camera.	The Ariane 6 rocket is visible in the image.	**30%**: The presence of the Ariane 6 rocket suggests a space-related mission, which aligns with the hypothesis about the Webb telescope. However, it does not confirm the specific purpose of capturing the first photons.
	The Webb telescope is shown in its deployment configuration.	**70%**: Seeing the Webb telescope in its deployment configuration strongly suggests that the technological development is related to the telescope, increasing the likelihood that the hypothesis is true.
	The Webb telescope's name is spelled out on the rocket's side.	**80%**: The Webb telescope's name on the rocket further confirms the association with the Webb telescope, making it quite likely that the purpose is to capture the first photons on Webb's primary camera.

Table 3: A sample question, the corresponding extracted evidence pieces, and resultant probability scoring rationale produced by BONSAI on the MultiVENT multiple choice task for an event in the "science and technology" domain.

modeling probabilities either penalize the system for comprehensive decompositions in the case of the Sec. approach, or fail to represent conditional dependencies in the case of the leaf score aggregation method used in these experiments. It is likely that the LLM judge ad-hoc models the dependencies while implicitly conditioning over tree size, resulting in a balanced aggregation approach.

Ambiguous Video Analysis

Task We evaluate BONSAI on MultiVENT (Sanders et al. 2023), a dataset of short to long-form videos depicting portions, but not the entirety, of various real-world current events. We scrape English news articles describing the events depicted in the dataset and use GPT-4o to generate various factual statements about these events. We then sample a set of 46 English document-video pairs using the MultiVENT 1.0 release, cluster a set of five statements about distractor events that are most semantically similar to the correct event statements (using a cross-encoder trained on MS-MARCO), and use these clusters as multiple-choice questions about the video content. We first solicit high-performing human crowdsource annotators to watch the videos and complete the QA task with two-way redundancy. We evaluate BONSAI's ability to identify the correct statement about each video, and use mean accuracy as the evaluation metric. We compare the system's performance against state-of-the-art video understanding models (with similar vision backbone size) and human performance (debatable vision backbone size). Human performance is reported via mean score and via the best individual annotator, who annotated all videos.

Setup We replicate the setup described in Sec. . For audio, we peform ASR with Whisper (Radford et al. 2022) and sample 6 "sentences" per window for evidence extraction. We demonstrate the value of "test time evidence extraction" by comparing multiple ways of sampling evidence. We first take the traditional approach used in Sec. , passing in a general question that each of the hypotheses attempt to answer. We compare this against re-sampling visual evidence using the leaf sub-claims as context (BONSAI$_{SL}$).

Results The results, shown in Table 2, illustrate BONSAI's ability to reason over vision-centric data where audio provides limited additional information, compared to a benchmark like TVQA where dialogue-only performance can reach over 44% (Sanders, Weir, and Van Durme 2024). The results also illustrate the efficacy of test-time evidence scaling, boosting BONSAI performance by 13 points and matching state-of-the-art black-box video model performance (and notably outperforming earlier 2024 models) while simultaneously providing comprehensive and grounded reasoning traces. While BONSAI does not significantly outperform Qwen-2 in this experiment, it is notable that it matches performance *while producing a comprehensive reasoning trace*. Human annotators rated their answer certainty below 70% for over a quarter of their answers on the task. Sample scoring outputs from BONSAI alongside the relevant extracted evidence is included in Table 3.

Conclusion

We introduce a probabilistic reasoning tree generator for broadly adaptable and transparent reasoning that remains grounded in multimodal evidence. Through this system we demonstrate that probabilistic reasoning via an iterative algorithm enables robust reasoning over uncertainty that improves human alignment both for low-level and high-level tasks. Further, it demonstrates the power of leveraging contextually conditioned captions from multimodal data to enable powerful cross-modal reasoning. As a general purpose system, BONSAI trades optimized performance on specific domains for broad adaptability, and as future work we envision significant improvement across different tasks by introducing domain-specific modifications to system modules. In short, BONSAI showcases the exciting potential of transparent reasoning systems on complex, real-world challenges.

References

Angelopoulos, A. N.; Bates, S.; et al. 2023. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4): 494–591.

Bajaj, P.; Campos, D.; Craswell, N.; Deng, L.; Gao, J.; Liu, X.; Majumder, R.; McNamara, A.; Mitra, B.; Nguyen, T.; et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Bekoulis, G.; Papagiannopoulou, C.; and Deligiannis, N. 2021. A review on fact extraction and verification. *ACM Computing Surveys (CSUR)*, 55(1): 1–35.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Cao, Y.; Nair, A. M.; Eyimife, E.; Soofi, N. J.; Subbalakshmi, K. P.; II, J. R. W.; Basu, C.; and Shallcross, D. 2024. Can Large Language Models Detect Misinformation in Scientific News Reporting? arXiv:2402.14268.

Chen, J.; Kim, G.; Sriram, A.; Durrett, G.; and Choi, E. 2024a. Complex Claim Verification with Evidence Retrieved in the Wild. arXiv:2305.11859.

Chen, T.; Jiang, Z.; Poliak, A.; Sakaguchi, K.; and Van Durme, B. 2019. Uncertain natural language inference. *arXiv preprint arXiv:1909.03042*.

Chen, Z.; Zhou, Q.; Shen, Y.; Hong, Y.; Sun, Z.; Gutfreund, D.; and Gan, C. 2024b. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1254–1262.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Dalvi, B.; Jansen, P.; Tafjord, O.; Xie, Z.; Smith, H.; Pipatanangkura, L.; and Clark, P. 2021. Explaining answers with entailment trees. *arXiv preprint arXiv:2104.08661*.

Dammu, P. P. S.; Naidu, H.; Dewan, M.; Kim, Y.; Roosta, T.; Chadha, A.; and Shah, C. 2024. ClaimVer: Explainable claim-level verification and evidence attribution of text through knowledge graphs. *arXiv preprint arXiv:2403.09724*.

Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J. S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Dmonte, A.; Oruche, R.; Zampieri, M.; Calyam, P.; and Augenstein, I. 2024. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*.

Feng, Y.; Zhou, B.; Lin, W.; and Roth, D. 2024. Bird: A trustworthy bayesian inference framework for large language models. *arXiv preprint arXiv:2404.12494*.

Ganguly, D.; Iyengar, S.; Chaudhary, V.; and Kalyanaraman, S. 2024. Proof of thought: Neurosymbolic program synthesis allows robust and interpretable reasoning. *arXiv preprint arXiv:2409.17270*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; and Wang, H. 2024a. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.

Gao, Z.; Niu, B.; He, X.; Xu, H.; Liu, H.; Liu, A.; Hu, X.; and Wen, L. 2024b. Interpretable contrastive monte carlo tree search reasoning. *arXiv preprint arXiv:2410.01707*.

Gupta, A.; and Srikumar, V. 2021. X-FACT: A New Benchmark Dataset for Multilingual Fact Checking. arXiv:2106.09248.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jansen, P. A.; Wainwright, E.; Marmorstein, S.; and Morrison, C. T. 2018. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. *arXiv preprint arXiv:1802.03052*.

Jiang, Z.; Liu, A.; and Van Durme, B. 2025. Conformal Linguistic Calibration: Trading-off between Factuality and Specificity. *arXiv preprint arXiv:2502.19110*.

Kao, W.-Y.; and Yen, A.-Z. 2024. MAGIC: Multi-Argument Generation with Self-Refinement for Domain Generalization in Automatic Fact-Checking. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 10891–10902. Torino, Italia: ELRA and ICCL.

Lanham, T.; Chen, A.; Radhakrishnan, A.; Steiner, B.; Denison, C.; Hernandez, D.; Li, D.; Durmus, E.; Hubinger, E.; Kernion, J.; et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Li, X.; Chen, S.; Kapadia, R.; Ouyang, J.; and Zhang, F. 2024. Minimal Evidence Group Identification for Claim Verification. arXiv:2404.15588.

Li, Y.; Zhang, R.; and Liu, J. 2024. An enhanced prompt-based LLM reasoning scheme via knowledge graph-integrated collaboration. In *International Conference on Artificial Neural Networks*, 251–265. Springer.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Luo, L.; Li, Y.-F.; Haffari, G.; and Pan, S. 2023. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*.

Lyu, Q.; Havaldar, S.; Stein, A.; Zhang, L.; Rao, D.; Wong, E.; Apidianaki, M.; and Callison-Burch, C. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Mei, X.; Yang, L.; Jiang, Z.; Cai, X.; Gao, D.; Han, J.; and Pan, S. 2024. An inductive reasoning model based on interpretable logical rules over temporal knowledge graph. *Neural Networks*, 174: 106219.

Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Mo, S.; and Xin, M. 2024. Tree of uncertain thoughts reasoning for large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12742–12746. IEEE.

Mohri, C.; and Hashimoto, T. 2024. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*.

Ozturkler, B.; Malkin, N.; Wang, Z.; and Jojic, N. 2022. Thinksum: Probabilistic reasoning over sets using large language models. *arXiv preprint arXiv:2210.01293*.

Paruchuri, A.; Garrison, J.; Liao, S.; Hernandez, J.; Sunshine, J.; Althoff, T.; Liu, X.; and McDuff, D. 2024. What are the odds? language models are capable of probabilistic reasoning. *arXiv preprint arXiv:2406.12830*.

Piriyakulkij, T.; Langenfeld, C.; Le, T. A.; and Ellis, K. 2024. Doing experiments and revising rules with natural language and probabilistic reasoning. *Advances in Neural Information Processing Systems*, 37: 53102–53137.

Press, O.; Hochlehnert, A.; Prabhu, A.; Udandarao, V.; Press, O.; and Bethge, M. 2024. CiteME: Can Language Models Accurately Cite Scientific Claims? arXiv:2407.12861.

Qiu, L.; Sha, F.; Allen, K. R.; Kim, Y.; Linzen, T.; and van Steenkiste, S. 2024. Can Language Models Perform Implicit Bayesian Inference Over User Preference States? In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*.

Quach, V.; Fisch, A.; Schuster, T.; Yala, A.; Sohn, J. H.; Jaakkola, T. S.; and Barzilay, R. 2023. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356.

Razghandi, A.; Hosseini, S. M. H.; and Baghshah, M. S. 2025. CER: Confidence Enhanced Reasoning in LLMs. *arXiv preprint arXiv:2502.14634*.

Sanders, K.; Etter, D.; Kriz, R.; and Van Durme, B. 2023. Multivent: Multilingual videos of events and aligned natural text. *Advances in Neural Information Processing Systems*, 36: 51065–51079.

Sanders, K.; Kriz, R.; Liu, A.; and Van Durme, B. 2022. Ambiguous images with human judgments for robust visual event classification. *Advances in Neural Information Processing Systems*, 35: 2637–2650.

Sanders, K.; Weir, N.; and Van Durme, B. 2024. Tv-trees: Multimodal entailment trees for neuro-symbolic video reasoning. *arXiv preprint arXiv:2402.19467*.

Srikanth, N.; and Rudinger, R. 2025. NLI under the Microscope: What Atomic Hypothesis Decomposition Reveals. *arXiv preprint arXiv:2502.08080*.

Stengel-Eskin, E.; Hase, P.; and Bansal, M. 2024. Lacie: Listener-aware finetuning for confidence calibration in large language models. *arXiv preprint arXiv:2405.21028*.

Thorne, J.; Vlachos, A.; Cocarascu, O.; Christodoulopoulos, C.; and Mittal, A. 2018. The fact extraction and VERification (FEVER) shared task. *arXiv preprint arXiv:1811.10971*.

Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; and Manning, C. D. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Tversky, A.; and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157): 1124–1131.

Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020. Fact or Fiction: Verifying Scientific Claims. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7534–7550. Online: Association for Computational Linguistics.

Wanner, M.; Ebner, S.; Jiang, Z.; Dredze, M.; and Durme, B. V. 2024. A Closer Look at Claim Decomposition. arXiv:2403.11903.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Weir, N.; Clark, P.; and Van Durme, B. 2022. Nellie: A neuro-symbolic inference engine for grounded, compositional, and explainable reasoning. *arXiv preprint arXiv:2209.07662*.

Weller, O.; Marone, M.; Weir, N.; Lawrie, D.; Khashabi, D.; and Durme, B. V. 2024. "According to ...": Prompting Language Models Improves Quoting from Pre-Training Data. arXiv:2305.13252.

Xiao, Y.; Liang, P. P.; Bhatt, U.; Neiswanger, W.; Salakhutdinov, R.; and Morency, L.-P. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv preprint arXiv:2210.04714*.

Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; and Hooi, B. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Xu, S.; Pang, L.; Shen, H.; Cheng, X.; and Chua, T.-S. 2024. Search-in-the-Chain: Interactively Enhancing Large Language Models with Search for Knowledge-intensive Tasks. arXiv:2304.14732.

Yeo, W. J.; Satapathy, R.; Goh, R. S. M.; and Cambria, E. 2024. How interpretable are reasoning explanations from prompting large language models? *arXiv preprint arXiv:2402.11863*.

Zhao, E.; Awasthi, P.; and Gollapudi, S. 2025. Sample, Scrutinize and Scale: Effective Inference-Time Search by Scaling Verification. *arXiv preprint arXiv:2502.01839*.

Zur, A.; Kreiss, E.; D'Oosterlinck, K.; Potts, C.; and Geiger, A. 2024. Updating CLIP to Prefer Descriptions Over Captions. arXiv:2406.09458.