BEYOND CORRECTNESS: THE MAXIMAL SUPPORTED ANSWER

Anonymous authors
Paper under double-blind review

ABSTRACT

When answering from retrieval evidence, language models face a fundamental dilemma: an answer can be factually true in the real world yet more specific than the provided document warrants, or evidence-grounded yet needlessly vague. Standard exact-match correctness merges these distinct behaviors, failing to penalize unsupported speculation or reward warranted specificity.

We formalize this trade-off through the **Maximal Supported Answer (MSA)**: the most informative candidate that holds across all possible worlds compatible with the given evidence. Under standard denotational semantics, the MSA is unique and monotonically refines as evidence strengthens. To empirically evaluate this behavior, we introduce **MAXSUPPORT**, a diagnostic benchmark pairing queries with systematic evidence interventions—additions, deletions, and placebos—across five semantic spaces (time, geography, numbers, taxonomies, and source attribution).

Furthermore, we present **LATTICECAL**, a practical, risk-calibrated selection policy that operationalizes this target over candidate closures without oracle access. Evaluated across seven representative instruction-tuned models (spanning 8B dense baselines to 235B/284B sparse MoEs and a commercial anchor), LATTICECAL achieves **64.5% exact MSA** while reducing unsupported detail violations from **28.8% to 4.2%** under finite-sample risk bounds, without requiring fine-tuning of the underlying generator. Our paired evaluation cleanly separates unwarranted overclaiming from excessive conservatism, offering a principled framework for evidence-grounded generation.

1 INTRODUCTION

Consider a factual query whose ground-truth leaf answer is 1994. An incomplete reference document states only that the event occurred *"in the 1990s."* When queried with this document, a modern language model frequently answers *"1994."* While the response is historically accurate in the real world, the supplied evidence does not distinguish 1994 from any other year in that decade. The model has not reasoned from the evidence; it has hallucinated ungrounded specificity from its parametric memory.

Now consider the inverse failure: when the reference document explicitly names *"July 14, 1994,"* the model cautiously outputs *"in the late twentieth century."* This response is undeniably supported by the text, yet it leaves critical evidential information stranded.

Current evaluation paradigms conflate these two phenomena under a monolithic notion of accuracy: exact-string matching rewards lucky parametric guesses, while factual-precision benchmarks (e.g., FActScore) reward unhelpful vagueness. Reliable evidence-grounded generation demands two complementary objectives: **avoiding unsupported detail (safety)** while **retaining all specificity licensed by the source (utility)**.

This tension is not unique to temporal expressions; it recurs across geographical regions, numerical ranges, taxonomic classifications, and source attributions. In each domain, candidate claims form an ordered hierarchy of semantic commitments—from broad abstractions to granular specifics. Answering at a finer grain than the source licenses constitutes an **overclaim**; answering at a coarser grain than permitted constitutes an **underclaim**.

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

We conceptualize evidence as defining a *version space* of possible worlds—all world states that remain consistent with the provided text. A claim is *evidence-supported* if and only if it holds across every world in this version space. Among all supported candidates, there exists an optimal resolution: the **Maximal Supported Answer (MSA)**, defined as the most informative claim that remains universally supported.

Crucially, as valid evidence is added, the version space narrows, and the MSA can only remain fixed or advance to a strictly finer resolution. Conversely, removing decisive sentences forces the warranted target to retreat. This formulation decouples three concepts that standard benchmarks merge:

$$\text{World-True} \neq \text{Evidence-Supported} \neq \text{Maximally-Supported}$$

A response can be world-true yet completely unsupported, or supported yet strictly suboptimal in specificity.

To make this evidential responsiveness empirically auditable, we construct **MAXSUPPORT**, an exact diagnostic benchmark spanning five structured semantic spaces. Unlike static QA datasets where evidence license is confounded by model pretraining priors, MAXSUPPORT employs *paired evidence ladders*: we hold the query fixed while programmatically injecting evidence additions, evidence deletions, and length-matched placebo texts. This paired design provides a functional behavioral test of whether a model's output resolution shifts in lockstep with evidential license.

Finally, we operationalize this objective in **LATTICECAL**, a deployable framework that couples model-generated candidate closures with conformal risk calibration. Rather than relying on fragile prompt engineering, LATTICECAL freezes an evidence-support scorer and calibrates hierarchical selection thresholds to guarantee provable finite-sample risk bounds. Across seven representative instruction-tuned models—including open-weight models from the Llama, Qwen, and DeepSeek families deployed on local A100 infrastructure, alongside a commercial API anchor—LATTICECAL achieves a macro **64.5% exact MSA** while curbing support violations to **4.2%** (down from 28.8% in native generation).

In summary, our contributions are:

1. **The MSA Formulation:** We formalize evidential answer resolution as the maximal supported answer on a version space, proving its uniqueness and monotonic behavior under clean denotational conditions.
2. **The MAXSUPPORT Benchmark:** We develop a paired-intervention benchmark across five semantic domains that causally isolates overclaiming, underclaiming, and evidence responsiveness.
3. **LATTICECAL & Empirical Findings:** We propose a non-oracle, risk-calibrated selection policy and demonstrate on seven representative instruction-tuned models that structured conformal calibration resolves the accuracy-granularity dilemma without fine-tuning the generative model.

2 RELATED WORK

Fine-Grained Factuality and the Granularity Paradox. Evaluating the factual reliability of large language models has evolved from simple string matching to fine-grained claim verification. Frameworks such as FACTSCORE (Min et al., 2023), LongFact, and SAFE (Wei et al., 2024) decompose long-form generations into individual atomic propositions and benchmark them against reference corpora, while FELM (Chen et al., 2023) localizes factual inaccuracies at the token and segment level. Similarly, citation-grounding benchmarks examine whether generated claims are strictly entailed by retrieved passages (Gao et al., 2023).

However, existing factuality metrics possess a structural blind spot: a system can easily achieve near-perfect factual precision simply by generating evasive, uninformative platitudes. Recent studies on answer informativeness (Huang et al., 2023) and multi-granularity QA (Yona et al., 2024) highlight this trade-off, yet prior uncertainty-adaptive methods like DRAG (Yona et al., 2024) calibrate granularity primarily against the model's internal confidence rather than the external evidence provided. In real-world retrieval, confidence and evidential license are orthogonal: specificity should advance only when the reference evidence becomes genuinely more resolving. MAXSUPPORT explicitly untangles this knot by evaluating unsupported precision and needless vagueness against a single, evidence-conditioned target.

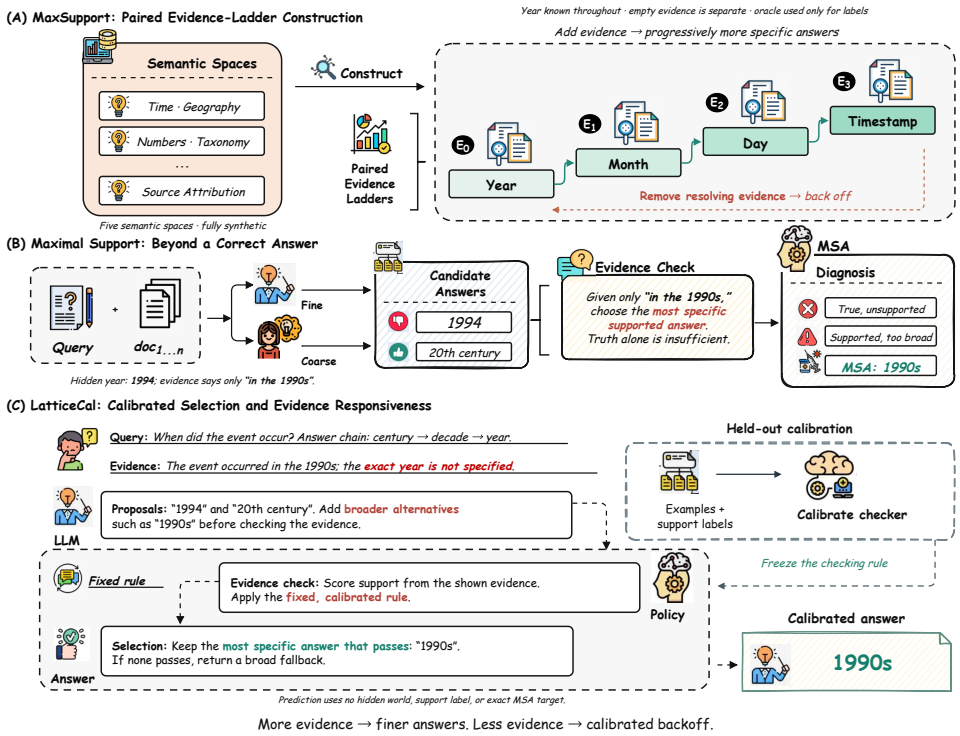

Figure 1: Overview of MAXSUPPORT and LATTICECAL. (A) The formal temporal hierarchy. (B–C) A teaching example distinguishes truth, support, and maximality; candidate scores feed held-out policy selection and calibrated risk control. The symbolic oracle supplies evaluation labels.

Selective Prediction and Hierarchical Abstention. When faced with uncertain inputs, selective prediction allows a model to abstain from answering to maintain high accuracy (Geifman & El-Yaniv, 2019). In semantic spaces with structured taxonomies, Hierarchical Selective Classification (HSC) expands the abstention action space: instead of a binary reject option, the model can safely retreat to an informative ancestor node (Goren et al., 2024). Conformal Linguistic Calibration models this trade-off as prediction sets over natural language descriptions (Jiang et al., 2025), while Selective Abstraction replaces uncertain fine-grained claims with coarser, verified generalizations in long-form generation (Goren et al., 2026).

Building upon these foundations, Compositional Selective Specificity (CSS) scores multiple renderings of varying granularity from a fixed draft (Huang et al., 2026). LATTICECAL operationalizes hierarchical selection over model-generated proposal closures, leveraging finite-sample conformal risk control (Angelopoulos et al., 2024) to formally guarantee that the deployed policy respects pre-declared error budgets across distinct semantic hierarchies.

Formal Antecedents in Knowledge Representation. Our denotational formulation of evidence-grounded answering draws deep lineage from classical database theory and description logics. In incomplete databases, the semantics of *certain answers* formalizes query answering as that which remains true across every possible completion of the database (Libkin, 2011). In description logics, the notions of *least common subsumers* (LCS) and *most specific concepts* (MSC) serve as representable concept summaries that capture commonalities without extraneous detail (Cohen et al., 1992; Baader, 2003). We bridge this formal lineage with modern open-ended generation, framing the optimal resolution as the most specific concept warranted by an evidence-induced version space and coupling it with causal intervention protocols.

Contextual Grounding and Evidence Responsiveness. FaithEval (Ming et al., 2025) and ConFiQA (Bi et al., 2025) test whether models follow supplied contexts under parametric conflict. These benchmarks primarily measure binary faithfulness—whether context is obeyed—rather than resolution warrant: how specifically the model is licensed to speak. Static test sets also fail to decouple

contextual reasoning from parametric memorization (Wan et al., 2024). By designing paired evidence ladders that track answer trajectories under controlled additions, deletions, and placebos, MAX-SUPPORT provides a functional behavioral test that isolates evidential responsiveness from static accuracy.

3 FROM EVIDENCE TO AN ANSWER RESOLUTION

Appendix Figure B.1 visualizes version-space contraction, the MSA boundary, support violations, and specificity regret.

INTUITIVE EXPOSITION AND ERROR DECOMPOSITION

To formalize the warranted resolution of an answer, consider a temporal query where the hidden world truth is 1994. The semantic candidate space consists of an uninformative root ($\perp$), the broad era *"the twentieth century"*, the decade *"the 1990s"*, and specific years such as *"1994"* and *"1995"*.

When an incomplete source document specifies only that the event transpired *"in the 1990s,"* all years within the range $[1990, 1999]$ remain viable possibilities. Under this state of knowledge:

- Outputting *"the twentieth century"* is factually grounded but leaves useful evidential license discarded.
- Outputting *"1994"* happens to match the hidden real-world truth, yet it constitutes an ungrounded speculation—it rules out nine other years that the evidence left completely open.
- Outputting *"the 1990s"* is both fully supported and maximally specific.

As illustrated in Figure B.1, reliable evidence-grounded answering cannot be characterized by a single scalar metric. An emitted response a under evidence E can fail in two orthogonal directions:

1. **Support Violation (SV):** The candidate rules out at least one world still compatible with the evidence, committing an unsupported overclaim:
$$\mathrm{SV}(a, E) = \mathbf{1}\{\mathcal{W}(E) \not\subseteq [\![a]\!]\}. \tag{1}$$
2. **Specificity Regret (SR):** The candidate is universally supported but remains unnecessarily coarse, incurring an information loss relative to the optimal target $m(E)$:
$$\mathrm{SR}(a, E) = I(m(E)) - I(a) \quad \text{(evaluated only when } \mathrm{SV}(a, E) = 0), \tag{2}$$
where $I(\cdot) \geq 0$ is a strictly increasing information-content functional over semantic refinement.

DENOTATIONAL ANSWER SYSTEMS AND THE MSA

We formally characterize this setup through denotational semantics. Let Ω denote a non-empty world space, representing the set of all mutually exclusive, complete states of the world. Let $\mathcal{A}$ be a finite set of answer candidates. Each candidate $a \in \mathcal{A}$ is mapped to its denotation $[\![a]\!] \subseteq \Omega$, denoting the set of possible worlds in which claim a holds true.

The candidate space is equipped with a semantic partial order $\preceq$ running from broad to specific:

$$a \preceq b \quad \iff \quad [\![b]\!] \subseteq [\![a]\!]. \tag{3}$$

We identify semantically equivalent surface forms with the same denotation and include a vacuous root element $\perp$ satisfying $[\![\perp]\!] = \Omega$.

Evidence text E induces a non-empty *version space* $\mathcal{W}(E) \subseteq \Omega$, representing the set of worlds that remain compatible with the statements in E. A candidate a is *evidence-supported* if the version space is a subset of its denotation, written as:

$$\mathrm{Supp}(E) = \{a \in \mathcal{A} : \mathcal{W}(E) \subseteq [\![a]\!]\}. \tag{4}$$

Because $\mathcal{W}(E) \subseteq \Omega = [\![\perp]\!]$, the root $\perp$ is always supported, ensuring $\mathrm{Supp}(E) \neq \emptyset$.

Support Completeness. In natural language, a query family may admit multiple fine-grained descriptors. To guarantee that a unique optimal answer exists rather than an incomparable set of maximal candidates, we declare the answer system *support-complete*: for every non-empty family of candidates $B \subseteq \mathcal{A}$ whose denotational intersection is non-empty, there exists a unique join $\bigvee B \in \mathcal{A}$ representing that exact intersection:

$$\left[\!\left[\bigvee B\right]\!\right] = \bigcap_{b \in B} [\![b]\!]. \tag{5}$$

Hierarchically nested concepts (e.g., standard taxonomies, temporal intervals, administrative regions) naturally fulfill this condition.

Theorem 1 (Existence, Uniqueness, and Evidence Monotonicity)**.** *In a finite support-complete answer system with a root $\bot$, every consistent evidence set E admits a unique greatest supported answer (which we designate as the **Maximal Supported Answer**, MSA), defined as the supremum of all supported candidates:*

$$m(E) = \bigvee_{a \in \mathrm{Supp}(E)} a. \tag{6}$$

Furthermore, if evidence E' is a valid refinement of E such that $\mathcal{W}(E') \subseteq \mathcal{W}(E)$, then the MSA behaves monotonically:

$$m(E) \preceq m(E'). \tag{7}$$

Proof Sketch. Support completeness represents the intersection of all supported denotations, which still contains $\mathcal{W}(E)$ and is therefore the unique greatest supported answer. Refinement gives $\mathrm{Supp}(E) \subseteq \mathrm{Supp}(E')$, hence $m(E) \preceq m(E')$; Appendix A provides the complete proof. □

Theorem 1 provides the bedrock for our diagnostic evaluation:

- When a **resolving clue** is introduced, the true target must monotonically advance ($m(E) \prec m(E')$).
- When that clue is **deleted**, the target must reversibly retreat ($m(E') \succ m(E)$).
- When a **length-matched placebo** is appended, the target must remain invariant ($m(E) = m(E \cup P)$).

Statistical Risk Formulation and Hierarchical Dominance

In realistic deployment, the true version space $\mathcal{W}(E)$ and denotational targets are withheld from the model. A practical answering policy $\pi(x, E)$ must predict a candidate from the text alone. To control the deployment risk of ungrounded speculation, we formulate policy evaluation under finite-sample conformal risk control.

For any complete selection policy π, we define its *selected-output support risk* over the data distribution $\mathcal{D}$ as:

$$R_{\mathcal{D}}(\pi) = \mathbb{E}_{(x,E)\sim\mathcal{D}}\left[\mathrm{SV}(\pi(x,E),E)\right] = \Pr_{(x,E)\sim\mathcal{D}}\left\{\mathcal{W}(E) \not\subseteq [\![\pi(x,E)]\!]\right\}. \tag{8}$$

Why should an answering system support intermediate abstractions rather than merely deciding whether to output the most specific leaf or abstain to root? The theoretical justification lies in action-space dominance:

Proposition 1 (Value of Intermediate Supported Answers)**.** *Let $\mathcal{F}_H(E) = \mathrm{Supp}(E)$ be the hierarchical action space, and let $\mathcal{F}_F(E) = \mathrm{Supp}(E) \cap (\mathrm{Leaves}(\mathcal{H}) \cup \{\bot\})$ be the flat (leaf-or-root) selective action space. Under exact support, the most informative hierarchical candidate weakly dominates the flat candidate:*

$$\max_{a \in \mathcal{F}_H(E)} I(a) \;\geq\; \max_{a \in \mathcal{F}_F(E)} I(a). \tag{9}$$

The gain is strictly positive whenever no leaf is supported, yet at least one intermediate ancestor node is licensed by the evidence.

The proof and strict-gain cases appear in Appendix A.

4 MAXSUPPORT AND LATTICECAL

EXACT PAIRED EVIDENCE LADDERS

MAXSUPPORT instantiates the MSA target across five structured semantic spaces: time, geography, numerical resolution, entity taxonomies, and source attribution (detailed in Table 3). Each query family fixes an underlying world fact (leaf ℓ) and constructs an evidence ladder consisting of an empty baseline condition followed by four progressively resolving evidence conditions E_0 through E_3, licensing target levels h_0 through h_3. The empty condition formalizes zero evidence, where the only supported action is the uninformative root $\perp$. A symbolic world model tracks the induced version space $\mathcal{W}(E)$ and computes the exact ground-truth MSA at each step.

The Role of Paired Boundary Facts. A key structural design of MAXSUPPORT is the systematic inclusion of *boundary facts*. In realistic language comprehension, establishing that an answer is supported requires verifying that alternative siblings are strictly excluded. For example, a document stating that an event occurred *"between 1990 and 1999"* supports the decade, but it does not by itself prove that a specific year is unresolved; it is the presence of an open boundary that keeps multiple child worlds live.

At rung r, positive ladders provide a resolving clue P_r licensing h_r alongside an implicit or explicit boundary clue B_r ensuring that no descendant h_{r+1} is uniquely entailed. Boundary clues include open descriptions such as *"sometime during the 1990s"* rather than relying on fixed negation keywords. Advancing to rung $r+1$ introduces a resolving clue while eliminating sibling decoys. To decouple contextual reasoning from superficial token overlap, we incorporate length-matched placebo texts that alter surface phrasing without contracting the version space. Furthermore, negative variants inject direct contradictions, sibling decoys, missing qualifiers, false presuppositions, or source mismatches. All query families, world entities, templates, and distractors are partitioned disjointly across splits to prevent ladder memorization between calibration and test.

Deterministic Oracle Role. The symbolic oracle performs two strictly bounded operations: constructing the gold MSA target and mapping parsed candidate text to its semantic denotation. Canonical model outputs are parsed using deterministic grammars; for unmatched surface expressions, a blinded equivalence verifier maps the string to the nearest declared candidate node without ever viewing the input evidence or judging support. The oracle is exclusively an evaluation artifact and is completely withheld during inference.

LATTICECAL: CALIBRATED HIERARCHICAL INFERENCE

LATTICECAL operationalizes evidence-conditioned answering without requiring oracle world access. Given an input query x, context E, and a declared hierarchy $\mathcal{H}$, the framework executes through five modular stages, summarized in Algorithm 1 and visualized in Figure D.1:

1. **Proposal Generation:** A frozen base generator $G(x, E)$ proposes an initial candidate set C_0 containing up to four distinct surface answers.
2. **Lattice Ancestor Closure:** To avoid overly conservative leaf-or-root abstention, LATTICECAL constructs the candidate closure by augmenting C_0 with all informative taxonomic ancestors in $\mathcal{H}$, plus the root fallback:
$$C(x, E) = C_0 \cup \text{Ancestors}_{\mathcal{H}}(C_0) \cup \{\perp\}. \tag{10}$$
3. **Support Scoring:** A trained cross-encoder $s_\theta(E, a) \in [0, 1]$ estimates the posterior probability that candidate a is unsupported given E. The empirical support confidence is $q_\theta(E, a) = 1 - s_\theta(E, a)$.
4. **Depth-Calibrated Filtering:** Rather than applying a uniform threshold across disparate semantic granularities, LATTICECAL applies a depth-specific threshold vector $\widehat{\boldsymbol{\tau}} = (\widehat{\tau}_0, \ldots, \widehat{\tau}_K)$. A candidate is admitted to the feasible set F if and only if its support confidence exceeds its depth threshold:
$$F = \left\{ a \in C(x, E) \setminus \{\perp\} : 1 - s_\theta(E, a) \geq \widehat{\tau}_{d(a)} \right\}. \tag{11}$$

Algorithm 1 LATTICECAL Prediction Policy

Require: Frozen generator G, support scorer s_θ, hierarchy $\mathcal{H}$, calibrated thresholds $\widehat{\boldsymbol{\tau}}$, tie order $\prec_0$
$C_0 \leftarrow G(x, E)$ ▷ Generate at most 4 initial candidate renderings
$C \leftarrow C_0 \cup \text{Ancestors}_{\mathcal{H}}(C_0) \cup \{\perp\}$ ▷ Expand to lattice ancestor closure
$F \leftarrow \left\{a \in C \setminus \{\perp\} : 1 - s_\theta(E, a) \geq \widehat{\tau}_{d(a)}\right\}$ ▷ Apply depth-stratified support thresholds
if $F = \emptyset$ **then**
  return $\perp$ ▷ Zero-risk uninformative fallback
end if
$D \leftarrow \arg\max_{a \in F} I(a)$ ▷ Filter candidates with maximal semantic depth
return the first member of D under canonical order $\prec_0$

5. **Information Maximization:** If F is empty, the policy safely retreats to $\perp$. Otherwise, it selects the candidate that maximizes semantic information content $I(a)$, resolving ties via a canonical order $\prec_0$.

Appendix Figure D.1 gives the complete proposal, closure, scoring, filtering, and fallback pipeline.

Family-level calibration. Support scoring uses a DeBERTa-v3-large cross-encoder trained on independent family splits and frozen before calibration. For each evaluated generator, we treat each of 250 held-out calibration families per semantic space as an independent Bernoulli trial and calibrate a model-specific depth-threshold vector $\widehat{\boldsymbol{\tau}}$ over a predeclared grid of 256 policies. Simultaneous one-sided Clopper–Pearson bounds use Bonferroni correction over spaces and policies. Since an item failure implies a failure of its containing family, $R_{\mathcal{D}}(\pi) \leq R_{\text{family}}(\pi) \leq \alpha$. The selected policy maximizes retained supported information subject to $\alpha = 0.05$ at confidence $1-\delta = 0.95$; the always-root policy is the safe fallback. Appendix D.2 gives the complete construction, information-access table, and the 21.6-point coverage versus 13.9-point selection-error decomposition.

5 EXPERIMENTS

Experimental Protocol and Scope. The evaluation protocol evaluates models across 5,000 endpoint families (800 regular and 200 negative families across each of the 5 semantic spaces), totaling 16,000 licensed ladder conditions and 1,250 calibration families. Long-form evaluation incorporates 240 composite multi-claim families evaluated across five length conditions $\{1, 3, 6, 9, 12\}$, yielding 1,200 total prompts.

All open-weight models (`Llama-3.1-8B-Instruct`, `Qwen2.5-32B-Instruct`, `DeepSeek-R1-Distill-Qwen-32B`, `Llama-3.3-70B-Instruct`, `Qwen3-235B-A22B-Instruct-FP8`, and `DeepSeek-V4-Flash`) are deployed locally on a single compute node housing $8\times$ NVIDIA A100-SXM4-80GB GPUs using vLLM and SGLang under tensor parallelism. We include `GPT-4o` via official APIs as a commercial reference anchor. Model outputs are obtained via greedy decoding ($T = 0.0$, top-$p = 1.0$) with fixed random seeds.

Evaluation Metrics. Exact MSA requires the selected node to equal $m(E)$; support violation (SV) marks unsupported precision; and informative underclaim (U-NR) records a supported, non-root strict ancestor. Total Root includes both correct negative actions and conservative positive fallbacks, so it overlaps with Exact MSA. Specificity-aware utility uses the reported rule $\text{MSA} + 0.5\,\text{U-NR} - \text{SV}$; formal definitions appear in Appendix D.3.

Endpoint Support and Information Retention. Table 1 shows that native generation attains 47.6% MSA with 28.8% SV, while prompting improves these to 56.4% and 18.1%. Flat selection reaches 4.7% SV only by returning root on 55.3% of items, limiting MSA to 45.8% and utility to 41.1. In contrast, LATTICECAL achieves **64.5% MSA**, **4.2% SV**, and the best deployable utility (72.7), outperforming HSC (63.9% MSA) and CSS (63.1%). The Candidate-set Oracle places proposal-closure coverage at 78.4%.

Figure 2 displays normalized answer depth against support violation across the threshold sweep. Depth measures resolution even when an answer is unsupported, so it is interpreted jointly with risk rather than as a standalone accuracy measure.

Table 1: **Endpoint results across seven models (%).** Macro averages over 5,000 families. Total Root overlaps with Exact MSA; utility is MSA $+ 0.5 \times$ U-NR $-$ SV. Red, orange, and yellow mark the top-three deployable policies.

Policy	MSA ↑	SV ↓	U-NR ↓	Total Root	Safe Neg. ↑	Utility ↑
Evaluated Policies						
Closed-book (Native)	47.6%	28.8%	**11.1%**	14.6%	28.1%	24.4%
Evidence prompting	56.4%	18.1%	**18.8%**	10.3%	46.6%	47.7%
Flat selective	45.8%	**4.7%**	**0.0%**	55.3%	71.0%	41.1%
HSC-style selection	**63.9%**	6.4%	20.5%	15.5%	**78.3%**	**67.8%**
CSS-style backoff	**63.1%**	**4.8%**	24.1%	14.5%	**80.1%**	**70.4%**
LATTICECAL	**64.5%**	**4.2%**	24.8%	13.4%	**84.2%**	**72.7%**
Oracle Reference						
Candidate-set Oracle	*78.4%*	*0.0%*	*14.3%*	*15.3%*	*100.0%*	*85.6%*

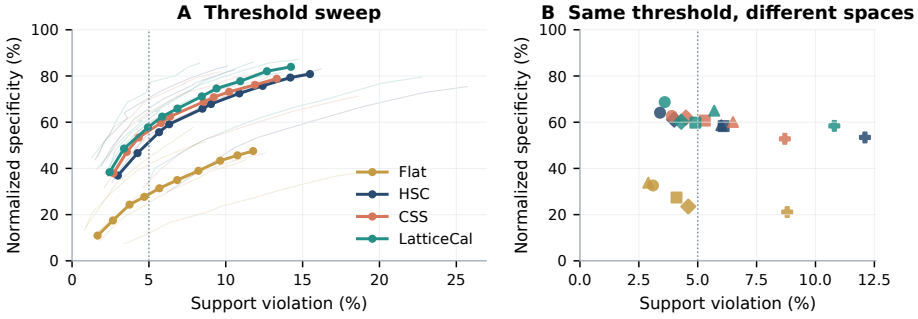

Figure 2: Risk–specificity tradeoffs. (A) Equal-space means (bold) and individual spaces (thin). (B) Spaces at threshold 0.80; shapes denote spaces and colors denote policies. Supplementary panels report all spaces.

Paired Evidence Responsiveness. A reliable system must demonstrate functional sensitivity to contextual changes. Table 2 reports pair-exact success rates where the model must correctly match gold targets under controlled evidence perturbations:

Under evidence addition ($E \rightarrow E^{+}$), standard evidence prompting achieves only 38.2% pair-exactness; models often anchor stubbornly to coarser historical outputs. Under deletion, prompting collapses to 34.5%, frequently retaining fine-grained assertions even after the decisive fact is expunged. In contrast, LATTICECAL achieves **56.8%** on additions and **55.4%** on deletions, indicating that policy-level calibration induces consistent, bidirectional evidence responsiveness. Placebo stability reaches 62.9%, demonstrating robustness against surface-level phrasing changes.

Semantic-Space Heterogeneity. Across individual domains, LATTICECAL achieves its highest exact resolution on time (72.8% MSA) and numerical ranges (68.5% MSA), where intermediate intervals exhibit uniform linear orderings. On entity taxonomy, HSC-style selection exhibits comparable performance (69.0% vs. 68.4% for LATTICECAL). Source attribution proves to be the most challenging space across all evaluated methods (LATTICECAL achieves 49.1% MSA), driven by the presence of partially incomparable revision DAGs.

Stress, Monotonicity, and Long-Answer Control. Across 35 model–space cells (Figure 3), LATTICECAL reduces mean monotonicity violations to 6.0% (Native: 26.3%; prompting: 16.7%) with a -7.0-point held-out shift. At nine claims, independent risk reaches 34.2%; simultaneous control under Theorem 5 reduces it to 8.0% (equal allocation) or 8.1% (value-weighted), with the latter raising retained key information from 76.5% to 84.1%. LATTICECAL also safely resolves 84.7% of contradictions and 88.0% of false presuppositions.

Table 2: **Paired evidence responsiveness (%).** Pair-exact rates require both outputs in an intervention pair to match their targets. Red, orange, and yellow mark the top-three rates.

Policy	Addition ↑	Deletion ↑	Placebo ↑
Evidence-conditioned Policies			
Evidence prompting	38.2%	34.5%	46.1%
HSC-style	**49.5%**	**48.0%**	**58.2%**
CSS-style	**51.2%**	**50.1%**	**60.5%**
LATTICECAL	**56.8%**	**55.4%**	**62.9%**

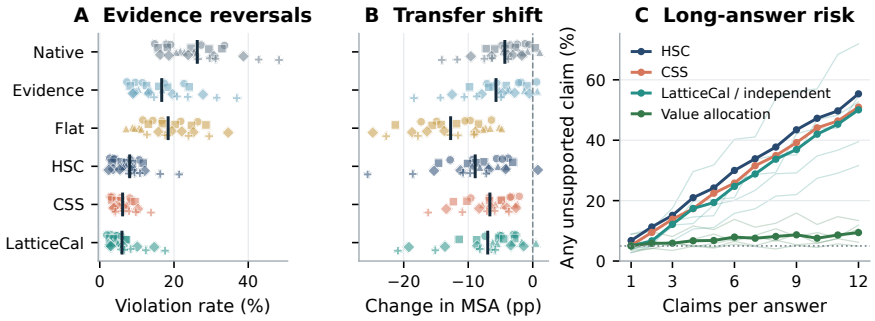

Figure 3: Stress and transfer results: monotonicity violations (A), held-out MSA shift (B), and document risk by claim count (C). Dots are model–space cells; thin LATTICECAL curves show individual spaces.

6 DISCUSSION AND LIMITATIONS

Three deployment boundaries remain. First, uniqueness requires a support-complete candidate space (Assumption A1); unconstrained open-domain DAGs therefore need upstream candidate induction (Appendix G). Second, paired interventions test functional responsiveness, not internal transformer mechanisms. Third, calibration assumes exchangeability: the -7.0-point cross-space drop favors domain-stratified calibration, while multi-sentence extensions additionally require joint coreference validation.

7 CONCLUSION

Reliable grounded answering must avoid unsupported detail while retaining warranted specificity. We formalize this target as the MSA, introduce MAXSUPPORT for paired evidence interventions across five semantic spaces, and develop LATTICECAL to combine proposal closures with finite-sample risk calibration. Across seven evaluated models, this reduces support violations from 28.8% to 4.2% while improving specificity-aware utility without requiring fine-tuning of the base generator.

8 AI USE STATEMENT

Generative AI tools assisted with literature organization, manuscript drafting, prose polishing, code generation, and figure design. The authors reviewed and verified the theoretical claims, empirical analyses, citations, code, and final presentation.

REPRODUCIBILITY STATEMENT

Complete proofs and metric formulations appear in Appendix A. Dataset construction rules, semantic spaces, parser protocols, calibration cohorts, and result tables are documented in the supplement. Section 5 reports the evaluated model roster, hardware, decoding settings, and fixed seeds.

REFERENCES

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

Franz Baader. Least common subsumers and most specific concepts in a description logic with existential restrictions and terminological cycles. In *IJCAI*, pp. 319–324, 2003. URL https://www.ijcai.org/Proceedings/03/Papers/047.pdf.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. Context-DPO: Aligning language models for context-faithfulness. In *Findings of ACL*, pp. 10280–10300, 2025. doi: 10.18653/v1/2025.findings-acl.536. URL https://aclanthology.org/2025.findings-acl.536/.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. FELM: Benchmarking factuality evaluation of large language models. In *NeurIPS Datasets and Benchmarks Track*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/8b8a7960d343e023a6a0afe37eee6022-Abstract-Datasets_and_Benchmarks.html.

William W. Cohen, Alexander Borgida, and Haym Hirsh. Computing least common subsumers in description logics. In *AAAI*, pp. 754–760, 1992. URL https://aaai.org/Papers/AAAI/1992/AAAI92-117.pdf.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *EMNLP*, pp. 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *ICML*, volume 97, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Shani Goren, Ido Galil, and Ran El-Yaniv. Hierarchical selective classification. In *NeurIPS*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/c8b100b376a7b338c84801b699935098-Abstract-Conference.html.

Shani Goren, Ido Galil, and Ran El-Yaniv. When should LLMs be less specific? selective abstraction for reliable long-form text generation. *arXiv preprint arXiv:2602.11908*, 2026. URL https://arxiv.org/abs/2602.11908.

Jie Huang, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. Can language models be specific? how? In *Findings of ACL*, 2023. URL https://aclanthology.org/2023.findings-acl.45/.

Tianyi Huang, Samuel Xu, Jason Tansong Dang, Samuel Yan, and Kimberley Yin. Answer only as precisely as justified: Calibrated claim-level specificity control for agentic systems. *arXiv preprint arXiv:2604.17487*, 2026. URL https://arxiv.org/abs/2604.17487.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. Conformal linguistic calibration: Trading-off between factuality and specificity. *arXiv preprint arXiv:2502.19110*, 2025. URL https://arxiv.org/abs/2502.19110.

Leonid Libkin. Incomplete information and certain answers in general data models. In *PODS*, pp. 59–70, 2011. doi: 10.1145/1989284.1989294.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*, 2023. URL https://aclanthology.org/2023.emnlp-main.741/.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. FaithEval: Can your language model stay faithful to context, even if "the moon is made of marshmallows"? In *ICLR*, 2025. URL https://openreview.net/forum?id=UeVx6L59fg.

Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing? In *ACL*, pp. 7468–7484, 2024. doi: 10.18653/v1/2024.acl-long.403. URL https://aclanthology.org/2024.acl-long.403/.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *NeurIPS*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/937ae0e83eb08d2cb8627fe1def8c751-Abstract-Conference.html.

Gal Yona, Roee Aharoni, and Mor Geva. Narrowing the knowledge evaluation gap: Open-domain question answering with multi-granularity answers. In *ACL*, pp. 6737–6751, 2024. doi: 10.18653/v1/2024.acl-long.365. URL https://aclanthology.org/2024.acl-long.365/.

This supplement contains complete proofs, counterexamples, the symbolic oracle construction, the full result tables, implementation details, and worked cases.

A Complete Formal Development

A.1 Denotational answer systems

Fix a nonempty world set Ω (finite or infinite). A denotational answer system consists of a finite candidate set $\mathcal{A}$, a partial order $\preceq$, and a map $[\![\cdot]\!] : \mathcal{A} \to 2^{\Omega}$ satisfying

$$a \preceq b \quad \Longleftrightarrow \quad [\![b]\!] \subseteq [\![a]\!].$$

The order runs from broad to specific. Surface forms with identical denotations are identified. The least element $\bot$ denotes Ω and represents the vacuous answer. An optional inconsistent element $\top$ denotes the empty set, but consistent evidence never supports it.

Evidence E induces a nonempty version space $\mathcal{W}(E) \subseteq \Omega$. We write $E \models a$ when $\mathcal{W}(E) \subseteq [\![a]\!]$, and define

$$\mathrm{Supp}(E) = \{a \in \mathcal{A} : E \models a\}.$$

All maximality statements are relative to this finite candidate language. Finiteness of the candidate set, rather than the world set, is used in the existence argument; interval denotations may have infinitely many members.

Support completeness. For every nonempty $B \subseteq \mathcal{A}$ whose denotational intersection is nonempty, assume that the join $\bigvee B$ exists and that

$$\Big[\!\Big[\bigvee B\Big]\!\Big] = \bigcap_{b \in B} [\![b]\!]. \tag{A1}$$

Nested denotations in a rooted taxonomy satisfy this condition. All supported candidates share the nonempty set $\mathcal{W}(E)$, so A1 applies to their join. An arbitrary collection of propositions may omit the required conjunctions.

Lemma 1 (Downward closure of support)**.** *If $a \preceq b$ and $E \models b$, then $E \models a$.*

Proof. The order gives $[\![b]\!] \subseteq [\![a]\!]$. Support of b gives $\mathcal{W}(E) \subseteq [\![b]\!]$. Transitivity of inclusion yields $\mathcal{W}(E) \subseteq [\![a]\!]$. □

Theorem 2 (Existence and uniqueness of maximal support)**.** *Under A1, every consistent evidence set has a unique greatest supported answer*

$$m(E) = \bigvee \mathrm{Supp}(E).$$

For every $a \in \mathrm{Supp}(E)$, $a \preceq m(E)$.

Proof. The root $\bot$ is supported because $\mathcal{W}(E) \subseteq \Omega = [\![\bot]\!]$, hence $\mathrm{Supp}(E)$ is nonempty. By A1 its join exists and

$$[\![m(E)]\!] = \bigcap_{a \in \mathrm{Supp}(E)} [\![a]\!].$$

For every supported a, the version space is contained in $[\![a]\!]$; it is therefore contained in their intersection. Thus $m(E)$ is itself supported. The defining property of a join gives $a \preceq m(E)$ for every supported a. If m and m' were both greatest, then $m \preceq m'$ and $m' \preceq m$; antisymmetry implies equality. □

Why A1 cannot be omitted. Let $\Omega = \{1, 2, 3\}$ and let the candidate set contain a root r with denotation Ω, a node a denoting $\{1, 2\}$, and an incomparable node b denoting $\{1, 3\}$. For $\mathcal{W}(E) = \{1\}$, both a and b are supported and maximal, but the candidate representing their conjunction is missing. There is no unique greatest supported answer. In such a task, the correct target is a set of maximal answers, or a single answer in an augmented language that represents the missing conjunction.

A.2 Evidence ladders and identification

Write $E \sqsubseteq E'$ when E' is at least as informative as E, meaning $\mathcal{W}(E') \subseteq \mathcal{W}(E)$.

Theorem 3 (Monotonicity under valid evidence refinement)**.** *Under A1, $E \sqsubseteq E'$ implies $m(E) \preceq m(E')$.*

Proof. If $a \in \mathrm{Supp}(E)$, then

$$\mathcal{W}(E') \subseteq \mathcal{W}(E) \subseteq [\![a]\!],$$

so $a \in \mathrm{Supp}(E')$. Hence $\mathrm{Supp}(E) \subseteq \mathrm{Supp}(E')$, and in particular $m(E)$ is supported at E'. Since $m(E')$ is greatest among the answers supported at E', we have $m(E) \preceq m(E')$. □

The theorem yields a directional benchmark prediction. For a nested positive ladder $E_0 \sqsubseteq \cdots \sqsubseteq E_K$, ideal resolution is nondecreasing. For the reverse deletion ladder, it is nonincreasing. This statement does not apply when added text changes source reliability, introduces a contradiction, or invokes nonmonotonic defaults. Those operations can enlarge or restructure the credible version space and require an explicit belief-revision rule.

Proposition 2 (Static correctness does not identify evidence use)**.** *Suppose an item has a true leaf ℓ and two evidence levels $E_0 \sqsubseteq E_1$ with $m(E_0) \prec \ell = m(E_1)$. The policies*

$$\pi_{\mathrm{responsive}}(E) = m(E), \qquad \pi_{\mathrm{fixed}}(E) = \ell$$

are observationally identical at E_1 but differ at E_0.

Proof. At E_1, both policies emit ℓ. At E_0, the responsive policy emits $m(E_0)$, whereas the fixed-leaf policy still emits ℓ. Since $m(E_0)$ is the greatest supported answer and $m(E_0) \prec \ell$, the leaf is unsupported at E_0. A test containing only E_1 cannot distinguish the policies; the paired condition can. □

The proposition identifies an output-level distinction. A paired test can separate these policies without identifying whether their behavior arose from memory, evidence reasoning, or a learned heuristic.

A.3 Support risk and the value of intermediate answers

For a consistent item define

$$Y(E, a) = \mathbf{1}\{\mathcal{W}(E) \not\subseteq [\![a]\!]\}.$$

For a complete policy π, its support risk is $R_{\mathcal{D}}(\pi) = \mathbb{E}_{(x,E)\sim\mathcal{D}} Y(E, \pi(x, E))$. If $a \preceq b$, then $Y(E, a) \leq Y(E, b)$ for the same item: when a broader answer excludes a compatible world, a narrower one excludes that world as well. The world-probability error $\Pr_{\omega\sim\mu_E}(\omega \notin [\![a]\!])$ is a different quantity. The support label is one whenever even a low-probability compatible world is excluded.

Let $I(a) \geq 0$ increase along meaningful refinements and let $I(\bot) = 0$. Write $\mathcal{F}_H(E) = \mathrm{Supp}(E)$ and $\mathcal{F}_F(E) = \mathrm{Supp}(E) \cap (\mathrm{Leaves}(H) \cup \{\bot\})$.

Theorem 4 (Strict gain under intermediate exact support)**.** *For each consistent item,*

$$\max_{a\in\mathcal{F}_H(E)} I(a) \geq \max_{a\in\mathcal{F}_F(E)} I(a).$$

The inequality is strict whenever no leaf is supported and some supported internal node has positive information value. If this event has positive probability, the expected gain is strict.

Proof. The flat feasible set is a subset of the hierarchical feasible set. On the specified event, the only flat action is $\bot$, with information zero, whereas the hierarchical set contains a supported node with positive information. The difference is nonnegative everywhere and positive on that event. □

For a marginal risk budget, inclusion of complete policy classes likewise gives weak dominance of their optimal expected utility at any common risk threshold. The pointwise strict statement above concerns exact support; it does not assume that a learned scorer recovers the optimal node.

Proposition 3 (Support-only scoring admits an uninformative optimum)**.** *An objective that minimizes support violations and assigns no value to information admits the always-root policy as an optimum on consistent items.*

Proof. The root denotes Ω and contains every nonempty version space, so the policy has zero violations, the minimum attainable value. □

Joint reporting of support violations, exact MSA, informative underclaims, and root frequency distinguishes this policy from informative supported answers.

A.4 Long answers and selected claims

Let a response contain J declared claim slots. Slot j has a finite answer space $\mathcal{A}_j$ and a predeclared finite class Π_j of complete resolution-selection policies. Calibration data C return an upper bound $U_j(\pi; C)$ for every $\pi \in \Pi_j$. Assume that there is an event G_j with $\Pr_C(G_j) \geq 1 - \delta_j$ such that, for every realized $C \in G_j$,

$$\Pr_X\{F_j(\pi, X) \mid C\} \leq U_j(\pi; C) \quad \text{for all } \pi \in \Pi_j, \tag{A2}$$

where F_j denotes an unsupported emitted claim on a fresh item. The simultaneous quantifier over complete policies is essential: separately valid bounds for fixed nodes do not automatically survive adaptive node selection.

Lemma 2 (Selection after simultaneous calibration)**.** *For $C \in G_j$, any rule selecting $\widehat{\pi}_j(C) \in \Pi_j$ only when $U_j(\widehat{\pi}_j; C) \leq \alpha_j$ satisfies $\Pr_X(F_j(\widehat{\pi}_j, X) \mid C) \leq \alpha_j$.*

Proof. For fixed $C \in G_j$, A2 holds simultaneously for every policy in Π_j and therefore for the selected policy. Combining the two inequalities proves the claim. □

Theorem 5 (Document-level support control)**.** *If A2 holds and $\sum_{j=1}^{J} \alpha_j \leq \alpha$, then*

$$\Pr_{C,X}\left(\bigcup_{j=1}^{J} F_j\right) \leq \alpha + \sum_{j=1}^{J} \delta_j.$$

If all claims share one joint calibration event of probability at least $1 - \delta$, the final term is δ.

Proof. Let $G = \bigcap_j G_j$. A union bound over calibration gives $\Pr_C(G^c) \leq \sum_j \delta_j$. Conditional on any fixed $C \in G$, the lemma gives $\Pr_X(F_j \mid C) \leq \alpha_j$ for every claim. A second union bound gives

$$\Pr_X\left(\bigcup_j F_j \mid C\right) \leq \sum_j \alpha_j \leq \alpha.$$

Integrating over $C \in G$ and accounting for G^c gives the result. The joint event variant is identical. □

The theorem is marginal over the calibration and deployment distributions. It does not guarantee a conditional error probability of α for every prompt. It also requires the deployment-time selection rule to be represented in the calibrated policy class. If budgets or claim selection depend on the prompt, their complete allocation rule must also belong to the simultaneously calibrated policy class; fixed-slot bounds alone do not justify an uncalibrated adaptive allocator.

Joint semantic coherence. Even exact local support composes only when all claims share one world and version space, with fixed referents and time scopes. Under that assumption, $\mathcal{W}(E) \subseteq [\![\widehat{a}_j]\!]$ for all j implies

$$\mathcal{W}(E) \subseteq \bigcap_{j=1}^{J} [\![\widehat{a}_j]\!].$$

Without the shared semantics, separately sourced claims can be individually supported and jointly incoherent.

Risk-budget allocation. Let $G_j(\alpha_j)$ be the expected information obtained from claim j at budget α_j. If every G_j is nondecreasing, concave, and differentiable, then the program

$$\max_{\alpha_j} \sum_j G_j(\alpha_j) \quad \text{s.t.} \quad \sum_j \alpha_j \le \alpha, \quad 0 \le \alpha_j \le \bar{\alpha}_j$$

is a standard concave resource-allocation problem. The KKT conditions give a multiplier λ such that every interior claim satisfies $G_j'(\alpha_j) = \lambda$; boundary claims satisfy the corresponding one-sided conditions. Equal allocation is optimal at an interior solution only when the marginal information gains coincide. Weighted allocation can improve the objective when those marginal gains differ. Establishing an empirical advantage requires measuring retained supported information as well as document risk.

B Counterexamples and Scope Boundaries

Inconsistent evidence. If trusted sources entail incompatible leaves, their raw version-space intersection is empty. Classical vacuous entailment would then support every answer. MaxSupport instead assigns a tagged conflict action, scored as root $\perp$; deployment systems require an explicit source-selection or belief-revision rule before applying the MSA construction.

Nonmonotonic revision. Theorem 3 applies only when new evidence contracts a nonempty version space. Retractions, corrections, and changes in source reliability can restructure that space and are evaluated as separate interventions.

Post-selection calibration. Selecting the most favorable candidate among many individually calibrated scores does not preserve their marginal guarantees. Our calibration therefore covers complete policies simultaneously, including proposal, threshold, tie, and fallback rules.

Incomparable answers and discourse coherence. When the candidate language omits a needed conjunction, several supported answers can be maximal and incomparable. For multi-claim responses, local support also fails to ensure that claims share a referent, event, or time scope. These cases require an antichain target or a joint denotational check.

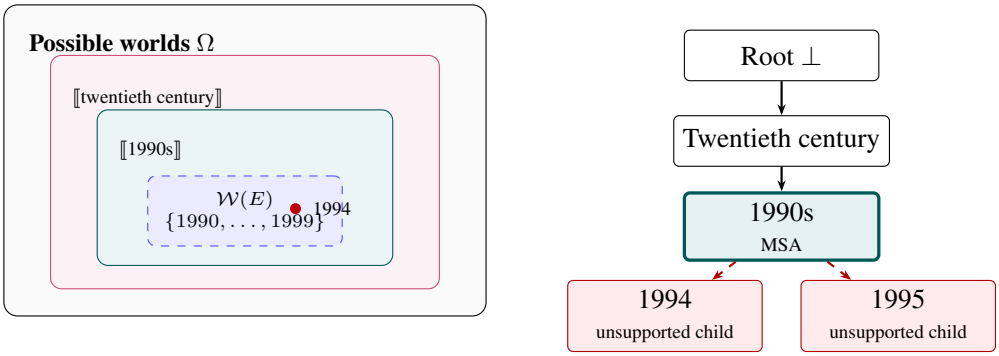

Figure B.1: Geometry of the maximal supported answer. Evidence restricts the live worlds to $\mathcal{W}(E)$. A candidate is supported exactly when its denotation contains that set. The 1990s is the finest supported node; choosing an ancestor loses licensed information, while choosing a year excludes live worlds even when it matches the hidden truth.

C SYMBOLIC ORACLE CONSTRUCTION

Table 3: Structured semantic spaces in MAXSUPPORT. Policies see textual evidence and the declared hierarchy; the oracle retains the world semantics.

Space	Leaf-to-root hierarchy	World semantics	Withheld boundary
Time	timestamp → day → month → year	UTC instants	finer instants remain live
Geography	venue → city → province → country	gazetteer cells	child location unresolved
Numbers	scalar → interval → bin → magnitude	bounded rationals	multiple child bins viable
Taxonomy	species → genus → family → order	tree membership	sibling leaves unresolved
Attribution	span → revision → work → archive	versioned records	revision or span withheld

Every regular family has four informative levels $h_0, \ldots, h_3$ and a separate empty-evidence condition whose target is $\perp$. At rung r, a support clue entails h_r and an implicit or explicit boundary clue ensures that no child is uniquely entailed. These clues include open descriptions such as *"sometime during the 1990s"* and do not depend on fixed negation keywords. Additions resolve a boundary, deletions restore it, and placebos change surface form without changing $\mathcal{W}(E)$.

Twenty percent of endpoint families are balanced across contradiction, sibling decoy, missing qualifier, false presupposition, and source mismatch. A deterministic grammar handles canonical outputs. Unmatched paraphrases are sent to a blinded equivalence verifier that sees neither evidence nor support labels; all final support and maximality decisions remain symbolic.

D METHODS AND EVALUATION PROTOCOL

D.1 CALIBRATED INFERENCE

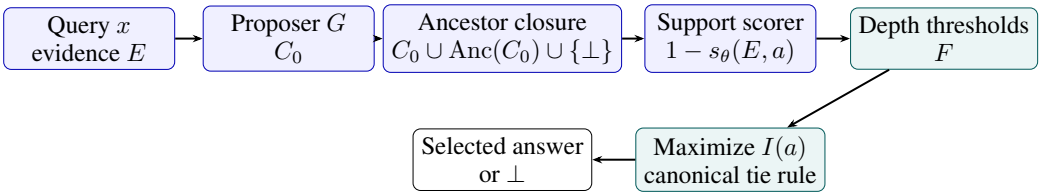

Figure D.1: LATTICECAL inference pipeline. Generation proposes candidate renderings; ancestor closure exposes safe intermediate actions; support scoring and depth-calibrated filtering produce a feasible set; the policy returns its most informative member or the root fallback.

D.2 FAMILY-LEVEL CALIBRATION

Support scoring uses a shared DeBERTa-v3-large cross-encoder trained on independent family splits and frozen before calibration. For each evaluated generator and semantic space, 250 calibration families are treated as independent Bernoulli trials, producing a model-specific threshold vector $\widehat{\boldsymbol{\tau}}$. The finite policy class contains 256 depth-threshold vectors plus the root fallback. For policy π and space d, let $k_{d,\pi}$ be the number of families containing any policy-level support failure. With $\eta = \delta/(5|\Pi|)$, the simultaneous one-sided bound is

$$U_d(\pi) = \mathrm{Beta}^{-1}(1 - \eta; k_{d,\pi} + 1, 250 - k_{d,\pi}).$$

Bonferroni correction covers all policies and spaces. The selected policy maximizes retained supported information subject to the weighted upper bound being at most $\alpha = .05$, with confidence $1 - \delta = .95$. Because an item failure implies a failure of its containing family, the construction also gives $R_{\mathcal{D}}(\pi) \le R_{\mathrm{family}}(\pi) \le \alpha$.

The data comprise 1,250 calibration families and 5,000 endpoint families. Four thousand regular families evaluated at four licensed rungs yield 16,000 ladder conditions. Long-answer evaluation uses 240 composite families at five claim lengths $\{1, 3, 6, 9, 12\}$, for 1,200 prompts.

D.3 METRICS

On consistent evidence, the target is the unique MSA $m(E)$. For tagged contradiction or presupposition-failure items with $\mathcal{W}(E) = \emptyset$, we define the extended target as $\widetilde{m}(E) = \bot$. Exact target resolution is $\mathbf{1}\{\widehat{a} = \widetilde{m}(E)\}$. Support violation is $\mathbf{1}\{\mathcal{W}(E) \not\subseteq [\![\widehat{a}]\!]\}$ on consistent items and $\mathbf{1}\{\widehat{a} \neq \bot\}$ on tagged negative items. Informative underclaim records a supported strict ancestor other than root, and Total Root records every $\bot$ output. Specificity-aware utility is

$$\text{MSA} + 0.5\,\text{U-NR} - \text{SV}.$$

Pair-exact evaluation requires both members of an addition, deletion, or placebo pair to match their evidence-conditioned targets. Families, rather than individual rungs, remain the resampling and calibration units.

E COMPLETE RESULTS

Table 4: Endpoint policy outcomes (%). Top-three deployable-policy values are shown in red, orange, and yellow, respectively.

Policy	MSA ↑	SV ↓	U-NR ↓	Total Root	Safe Neg. ↑	Utility ↑
Evaluated Policies						
Closed-book (Native)	47.6%	28.8%	**11.1%**	14.6%	28.1%	24.4%
Evidence prompting	56.4%	18.1%	**18.8%**	10.3%	46.6%	47.7%
Flat selective	45.8%	**4.7%**	**0.0%**	55.3%	71.0%	41.1%
HSC-style	**63.9%**	6.4%	20.5%	15.5%	**78.3%**	**67.8%**
CSS-style	**63.1%**	**4.8%**	24.1%	14.5%	**80.1%**	**70.4%**
LATTICECAL	**64.5%**	**4.2%**	24.8%	13.4%	**84.2%**	**72.7%**
Oracle Reference						
Candidate-set Oracle	*78.4%*	*0.0%*	*14.3%*	*15.3%*	*100.0%*	*85.6%*

The Candidate-set Oracle covers 78.4% of targets. Relative to LATTICECAL, the 35.5-point total error decomposes into a 21.6-point coverage gap and a 13.9-point selection gap. Flat selection obtains low SV primarily through its 55.3% root rate.

Table 5: Seven-model endpoint breakdown (%). Top-three values are shown in red, orange, and yellow, respectively.

Model	Native MSA ↑	Native SV ↓	LATTICECAL MSA ↑	LATTICECAL SV ↓
Open-weight Models				
Llama-3.1-8B-Instruct	40.5%	36.2%	56.4%	5.6%
Qwen2.5-32B-Instruct	44.2%	31.8%	61.5%	4.8%
DeepSeek-R1-Distill-Qwen-32B	47.8%	27.5%	64.2%	4.1%
Llama-3.3-70B-Instruct	46.5%	29.4%	63.8%	4.3%
Qwen3-235B-A22B-Instruct-FP8	**49.8%**	**25.6%**	**67.2%**	**3.8%**
DeepSeek-V4-Flash	**51.4%**	**24.2%**	**68.9%**	**3.4%**
Proprietary Models				
OpenAI GPT-4o	**53.0%**	**26.9%**	**69.5%**	**3.4%**
Macro Average	**47.6%**	**28.8%**	**64.5%**	**4.2%**

Table 6: Mean selected depth across evidence conditions. Root is -1 and the four informative levels are 0 through 3.

Policy	Empty	E_0	E_1	E_2	E_3
Mean Selected Depth					
Closed-book (Native)	2.8	2.8	2.8	2.8	2.8
Evidence prompting	1.5	1.8	2.2	2.7	2.9
HSC-style	0.5	1.0	1.7	2.5	2.9
CSS-style	0.3	0.9	1.6	2.6	2.9
LATTICECAL	0.2	0.8	1.5	2.6	2.9

At nine claims, independent thresholds yield 34.2% document risk. Equal allocation yields 8.0%, while value-weighted allocation yields 8.1% and retains 84.1% of key information. Negative-case safe resolution for LATTICECAL is 84.7% on contradiction, 85.5% on sibling decoys, 89.4% on missing qualifiers, 88.0% on false presuppositions, and 73.4% on source mismatch.

F WORKED SEMANTIC EXAMPLES

If the hidden year is 1994 but the document licenses only the 1990s, answering "1994" is world-true but unsupported, answering "the twentieth century" is supported but unnecessarily coarse, and answering "the 1990s" is the MSA. For provenance, evidence that identifies a document family but not a revision licenses the family node: a revision overclaims, while the repository root underclaims.

G EXTENDED DEPLOYMENT LIMITATIONS

Candidate-relative exactness. Open-domain systems must induce or retrieve a meaningful answer order. If the language omits useful conjunctions, a single MSA is not justified.

Calibration shift. Finite-sample guarantees apply to the declared family distribution. Production systems should maintain domain-specific calibration sets and audit coverage under template, hierarchy, and source shift.

Realization preservation. Selecting a safe node does not ensure that a later free-form realization stays within its denotation. A deployed system should constrain or reparse the final text to detect unsupported modifiers.