CERTIFIED COHERENT REASONING FOR LLMS VIA WEIGHTED MAXSAT AND BELIEF-REVISION GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) frequently produce mutually incompatible answers across sets of related questions, limiting reliability in domains that require globally consistent deduction, abduction, and theory maintenance. We formalize *coherent multi-query answering* as an optimization problem: an LLM supplies candidate answers with preference scores, logic supplies global feasibility constraints, and we select an answer set maximizing preference subject to satisfiability. Our main technical contribution is a *proof-carrying coherence decoder* that compiles the selection problem to weighted partial MaxSAT, outputs a globally consistent answer set, and emits certificates enabling independent verification of feasibility (and optionally optimality). We introduce solver-grounded coherence metrics, including the *coherence gap* (log-preference sacrificed to achieve global satisfiability) and *minimal-change repair distance* (belief-revision style). Theoretically, we give equivalences to distance-based belief revision (including a Dalal-style specialization), a complexity landscape (NP-hardness under severe restrictions), and a probabilistic extension as a KL projection onto the coherent support set. We also propose benchmark templates (pure symbolic and controlled natural language) in which the semantic map to logic is unambiguous, enabling clean measurement of global coherence, contradiction localization, and certified repair.

1 INTRODUCTION

LLMs have strong local fluency yet weak *global* logical reliability: across multiple related questions about a shared latent situation, they often return answers that cannot jointly hold. This is not merely a "contradiction rate" issue; the core difficulty is that global coherence is a *constrained theory selection* problem: the model produces preferences over many locally plausible statements, while consistency requires selecting a subset that admits a single model of the world.

This paper proposes a solver-grounded framework in which: (i) the LLM provides a structured *preference distribution* over candidate answers; (ii) logic provides *feasibility constraints* shared across questions; (iii) a MaxSAT optimizer selects the best globally consistent answer set; and (iv) we output *certificates* (proof-carrying reasoning) so third parties can verify coherence.

The key perspective is:

> *LLM outputs are preferences; logic is feasibility; MaxSAT is coherent decoding; certificates are trust.*

Contributions.

1. **Coherent decoding as constrained theory selection.** We define a multi-question answer selection problem with global satisfiability constraints and show how to compile it to weighted partial MaxSAT.
2. **Coherence geometry + metrics.** We introduce solver-native coherence measures: coherence gap, minimal-change distance, and MUS/MCS-derived conflict centrality.

3. **Proof-carrying coherence.** We formalize feasibility certificates (models) and discuss optional optimality proof logging, enabling independently checkable coherence.
4. **Theory.** We establish NP-hardness (even with $K = 2$ and clause-level candidates), connect to distance-based belief revision, and provide a probabilistic extension as an I-projection onto coherent supports.
5. **Benchmarks.** We propose benchmark layers that avoid "semantic parsing ambiguity" by construction.

2 Problem Setting: Coherent Multi-Query Answer Selection

2.1 Questions, candidates, preferences, and semantics

Let $Q = \{q_1, \ldots, q_n\}$ be questions about a shared latent world. For each i, an LLM produces K candidate answers

$$a_{i1}, \ldots, a_{iK} \quad \text{with scores} \quad w_{ij} := \log P_\theta(a_{ij} \mid q_i, c),$$

where c is shared context.

Each candidate is mapped to a formula in a chosen logic fragment $\mathcal{L}$:

$$\Phi(q_i, a_{ij}) =: \varphi_{ij} \in \mathcal{L}.$$

Let $B \in \mathcal{L}$ be a background theory encoding shared constraints (domain axioms, extracted facts, task rules, etc.).

Assumption 2.1 (Semantic determinism layer)**.** For benchmarking and certification, we assume Φ is deterministic and produces formulas whose satisfiability is decidable in the chosen fragment $\mathcal{L}$. In practice, we recommend either (i) purely symbolic tasks where Φ is trivial, or (ii) controlled natural language with a deterministic compiler.

2.2 Selection variables and coherent theory

Let $s \in \{0,1\}^{n \times K}$ be a selection matrix with the constraint

$$\sum_{j=1}^{K} s_{ij} = 1 \quad \forall i \in [n]. \tag{1}$$

Define the selected theory:

$$T(s) \;:=\; B \;\cup\; \{\varphi_{ij} : s_{ij} = 1\}.$$

Definition 2.2 (Optimal coherent answer set)**.** The *optimal coherent selection* is the solution of:

$$\max_{s \in \{0,1\}^{n \times K}} \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} s_{ij} \quad \text{s.t.} \quad \mathrm{SAT}\big(T(s)\big). \tag{2}$$

This is *global coherence by construction*: feasibility is satisfiability of the joint theory.

3 Compilation to Weighted Partial MaxSAT

We now give a concrete compilation to WCNF so that any weighted partial MaxSAT solver can compute the coherent selection.

3.1 CNF normalization and gating

For each candidate formula φ_{ij}, we obtain an equisatisfiable CNF $\mathrm{CNF}(\varphi_{ij})$ (e.g., Tseitin-style linear-size transformation). Introduce selector variables $z_{ij} \in \{0,1\}$ indicating "candidate (i,j) is chosen." Enforce selection via OneHot constraints on each row $(z_{i1}, \ldots, z_{iK})$.

To ensure only the chosen candidate contributes constraints, gate each formula:

$$z_{ij} \to \varphi_{ij}.$$

In CNF, gating can be applied clausewise: if $\mathrm{CNF}(\varphi_{ij}) = \bigwedge_t C_{ijt}$, add hard clauses

$$(\neg z_{ij} \;\vee\; C_{ijt}) \quad \forall t.$$

Exactly-one encoding. We encode $\mathrm{OneHot}(z_{i1}, \ldots, z_{iK})$ as:

$$(z_{i1} \vee \cdots \vee z_{iK}) \quad \wedge \quad \bigwedge_{j<\ell} (\neg z_{ij} \vee \neg z_{i\ell})$$

(pairwise), or use a linear-size sequential-counter encoding for scalability (Sinz, 2005).

3.2 Weights and invariances

MaxSAT solvers typically require nonnegative integer weights. We therefore apply a per-question shift:

Definition 3.1 (Weight shift)**.** Let $\alpha_i := \min_j w_{ij}$. Define shifted weights

$$\mathrm{Shift}(w_{ij}) := w_{ij} - \alpha_i \geq 0.$$

This shift preserves argmax selections because $\sum_i \alpha_i$ is constant under equation 1.

3.3 WCNF instance

Construct a weighted partial MaxSAT instance:

- **Hard clauses**: $\mathrm{CNF}(B)$, all OneHot constraints, and all gating clauses $(\neg z_{ij} \vee C_{ijt})$.
- **Soft clauses**: unit clauses (z_{ij}) with weight $\mathrm{Shift}(w_{ij})$.

Theorem 3.2 (Equivalence to MaxSAT decoding)**.** *Let $\mathcal{I}$ be the above* WCNF *instance. Any optimal MaxSAT assignment $\mu^\star$ induces a selection $s^\star_{ij} = \mu^\star(z_{ij})$ that solves equation 2 (with shifted weights), and conversely any optimal coherent selection induces an optimal MaxSAT assignment.*

Proof sketch. Hard clauses enforce OneHot and gating, hence under any satisfying assignment exactly one z_{ij} per i is true and the chosen candidates' formulas hold jointly with B. The soft objective is precisely $\sum_{ij} \mathrm{Shift}(w_{ij}) z_{ij}$, so maximizing soft weight corresponds to maximizing preference among satisfiable selections. Full proof in Appendix **??**. □

4 Coherence Geometry and Solver-Grounded Metrics

We introduce metrics defined directly from equation 2 and its MaxSAT structure.

4.1 Coherence gap

Define the local (independent) score

$$W_{\mathrm{loc}} := \sum_{i=1}^{n} \max_j w_{ij},$$

and the coherent optimum

$$W^\star := \max_{s:\mathrm{SAT}(T(s))} \sum_{i,j} w_{ij} s_{ij}.$$

Definition 4.1 (Coherence gap)**.**

$$\mathrm{CG} := W_{\mathrm{loc}} - W^\star \quad \geq 0.$$

Proposition 4.2 (Basic properties)**.** $\mathrm{CG} = 0$ *iff the independently best answers are jointly satisfiable with B. Moreover,* CG *is invariant under per-question weight shifts.*

4.2 Minimal-change repair distance

Let s^0 be a baseline selection (e.g., top-1 per question). Define:

$$\mathrm{MC}(s^0) := \min_{s:\mathrm{SAT}(T(s))} \|s - s^0\|_0, \tag{3}$$

i.e., the minimal number of answers that must change to restore satisfiability.

Remark 4.3. $\mathrm{MC}(s^0)$ is a belief-revision distance on the discrete selection space. It can be computed by a second MaxSAT instance where deviations from s^0 incur unit penalties (Appendix **??**).

4.3 MUS/MCS STRUCTURE AND CONFLICT CENTRALITY

When a proposed selection is inconsistent, we want *structure*, not just a binary "contradiction." Let $\mathcal{F}$ be a set of clauses (or clause groups) induced by B and selected candidates.

Definition 4.4 (MUS/MCS (informal))**.** A *minimal unsatisfiable subset* (MUS) is a subset of $\mathcal{F}$ that is unsatisfiable, but becomes satisfiable if any element is removed. A *minimal correction set* (MCS) is a minimal set of elements whose removal makes $\mathcal{F}$ satisfiable.

Definition 4.5 (Conflict centrality)**.** For a candidate (i, j), define centrality as the number of MUSes (under a chosen clause-grouping scheme) in which its gated constraints participate. This yields an explainable notion of "which answer causes contradictions."

5 THEORETICAL RESULTS

5.1 CONSISTENCY-BY-CONSTRUCTION

Theorem 5.1 (Feasibility guarantee)**.** *If the MaxSAT solver returns an assignment μ satisfying all hard clauses, then the induced theory $T(s)$ (where $s_{ij} = \mu(z_{ij})$) is satisfiable in $\mathcal{L}$; thus the selected answers are globally non-contradictory in $\mathcal{L}$.*

Proof sketch. Hard clauses include $\mathrm{CNF}(B)$ and gated $\mathrm{CNF}(\varphi_{ij})$ for each selected (i, j), hence any satisfying assignment is a model of B and all selected candidate formulas. Therefore, $T(s)$ is satisfiable. □

5.2 NP-HARDNESS UNDER SEVERE RESTRICTIONS

Theorem 5.2 (NP-hardness with $K = 2$ and clause candidates)**.** *The optimization problem equation 2 is NP-hard even when: (i) $K = 2$ for every question, (ii) $\mathcal{L}$ is propositional logic, and (iii) every φ_{ij} is a single CNF clause (or $\top$).*

Proof sketch. Reduce from weighted partial MaxSAT. For each soft clause C_t with weight w_t, create a question with two candidates: include C_t (weight w_t) or include $\top$ (weight 0). Let B encode the hard clauses. Any coherent selection corresponds to choosing a satisfiable subset of soft clauses with maximum total weight, matching MaxSAT. Full details in Appendix **??**. □

5.3 BELIEF-REVISION EQUIVALENCE (DISTANCE GEOMETRY)

We interpret selections as points in a discrete space (a product of simplices). Minimal-change repair becomes a distance-based revision operator.

Theorem 5.3 (Dalal-style specialization)**.** *Fix a baseline selection s^0. Suppose $\mathcal{L}$ is propositional and each question corresponds to choosing a literal assignment for a designated variable (i.e., $K = 2$ encodes x_i = True/False). Then minimizing $\mathrm{MC}(s^0)$ subject to satisfiability is equivalent to selecting a satisfying assignment of minimum Hamming distance from the baseline assignment (a Dalal-type revision step).*

Proof sketch. Under the literal-assignment encoding, $\|s - s^0\|_0$ equals Hamming distance between truth assignments. The feasible set is exactly the set of models of B. Thus equation 3 computes the nearest model in Hamming distance. □

5.4 PROBABILISTIC EXTENSION: KL PROJECTION ONTO THE COHERENT SUPPORT

Define an unconstrained Gibbs distribution over selections:

$$Q_\theta(s) \;\propto\; \exp\Big(\sum_{i,j} w_{ij} s_{ij}\Big), \quad \text{with } s \text{ satisfying equation 1.}$$

Let $\mathcal{C} := \{s : \mathrm{SAT}(T(s))\}$ be the coherent set.

Algorithm 1 Certified Coherent Decoding via Weighted Partial MaxSAT

Input: Questions $Q = \{q_i\}_{i=1}^{n}$, context c, candidates $\{a_{ij}\}$ with scores $\{w_{ij}\}$, background B, semantic map Φ, candidate budget K.
Output: Coherent answers $\{a_{ij^\star(i)}\}$ and certificate(s).
for $i = 1$ to n **do**
  for $j = 1$ to K **do**
    Compute $\varphi_{ij} \leftarrow \Phi(q_i, a_{ij})$.
    Compute $\mathrm{CNF}(\varphi_{ij})$ (e.g., Tseitin if needed).
  end for
end for
Introduce selector vars z_{ij} and encode $\mathrm{OneHot}(z_{i1}, \ldots, z_{iK})$ for all i.
Add hard clauses: $\mathrm{CNF}(B)$, all OneHot, and all gated clauses $(\neg z_{ij} \vee C_{ijt})$.
Add soft clauses: (z_{ij}) with weights $\mathrm{Shift}(w_{ij})$.
Run a weighted partial MaxSAT solver to obtain assignment $\mu^\star$ and (optional) proof log.
Decode selection: $j^\star(i) \leftarrow \{j : \mu^\star(z_{ij}) = 1\}$.
Emit: coherent answers $\{a_{ij^\star(i)}\}$, feasibility certificate $\mu^\star$, and optional optimality proof.

Theorem 5.4 (I-projection onto coherent support)**.** *Assume $Q_\theta(\mathcal{C}) > 0$. Then*

$$P^\star \;=\; \operatorname*{arg\,min}_{P:\mathrm{supp}(P)\subseteq\mathcal{C}} D_{\mathrm{KL}}(P\|Q_\theta)$$

is the conditional distribution $P^\star(s) = Q_\theta(s \mid s \in \mathcal{C})$.

Proof. This is the standard property of KL minimization under a hard support constraint: the minimizer preserves Q_θ's relative probabilities on the allowed set and renormalizes. A full derivation is in Appendix **??**. □

Remark 5.5 (Computation)**.** Computing $Q_\theta(\mathcal{C})$ is a constrained partition function (a form of weighted model counting), which is #P-hard in general (Valiant, 1979). This yields a clean "theory section": exact inference is intractable; approximate sampling / variational bounds become principled approximations.

6 Proof-Carrying Coherent Answers

6.1 Feasibility certificates (models)

Given a selected answer set $s^\star$, the solver can output a satisfying assignment μ over all propositional variables in the compiled instance, including the latent world variables and z selectors. Verification is polynomial:

1. check OneHot constraints,
2. check gated clauses, and
3. check $\mathrm{CNF}(B)$.

6.2 Optional optimality certificates (proof logging)

Beyond feasibility, one may want *optimality* proofs for MaxSAT. SAT has mature unsatisfiability proof formats (e.g., DRAT (Heule, 2016)). MaxSAT proof logging has recently become practical in solver ecosystems (**?**). Our framework is compatible with emitting and checking such logs, enabling *verifiable optimal coherence* rather than "trust the solver".

7 Algorithm

8 Benchmark Design for Clean Coherence Evaluation

To avoid conflating coherence with semantic parsing noise, we propose two benchmark layers.

8.1 Layer 1: purely symbolic multi-question coherence

Generate SAT-derived worlds over variables $x_1, \ldots, x_m$. Questions query values of variables, derived literals, or consequences. Candidates are $\{\texttt{True}, \texttt{False}\}$, so Φ is trivial. This layer isolates *global coherence* and solver behavior.

8.2 Layer 2: controlled natural language with deterministic compilation

Template symbolic constraints into controlled English (deterministic mapping to CNF). This tests the LLM's ability to propose correct candidates under linguistic surface variation, while preserving unambiguous evaluation.

8.3 Evaluation protocol

Report: (i) CG and $\mathrm{MC}(s^0)$, (ii) success rate of producing satisfiable global answer sets, (iii) MUS/MCS statistics for contradiction structure, and (iv) tradeoffs between coherence and preference loss.

9 Discussion and Limitations

Semantic map Φ. Our guarantees hold in the chosen fragment $\mathcal{L}$ *after* compilation. If Φ is noisy, the system certifies coherence of the compiled meaning, not necessarily of informal natural language semantics. This motivates controlled-language benchmarks and explicit parser uncertainty modeling.

Expressivity vs decidability. Richer $\mathcal{L}$ (e.g., SMT) increases modeling power but changes solver backends and proof formats. The MaxSAT compilation remains valuable whenever logical constraints can be reduced to (pseudo-)Boolean form.

Scalability. Modern MaxSAT solvers are competitive on large instances (**?**); nonetheless, worst-case complexity remains. Practically, one can cap K, use incremental solving, or switch to core-guided approximate variants.

10 Conclusion

We presented a theory-first framework for *certified coherent reasoning* across multiple related LLM questions. By posing answer selection as a maximum-weight satisfiable theory problem, compiling to weighted partial MaxSAT, and emitting certificates, we obtain coherence by construction and verifiability by design. The framework yields mathematically sharp coherence metrics, clean connections to belief revision, and a principled probabilistic extension via KL projection onto coherent supports. This provides a rigorous spine for future work on solver-aided LLM deduction, contradiction avoidance across dialogues, and trustworthy reasoning pipelines.

References

Mukesh Dalal. Investigations into a theory of knowledge base revision. In *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-88)*, pp. 475–479, 1988.

Marijn J. H. Heule. The drat format and drat-trim checker. *CoRR*, abs/1610.06229, 2016.

Hannes Ihalainen, Andy Oertel, Yong Kiam Tan, Jeremias Berg, Matti Järvisalo, Magnus O. Myreen, and Jakob Nordström. Certified maxsat preprocessing. In *Automated Reasoning – IJCAR 2024*, volume 14739 of *Lecture Notes in Computer Science*, pp. 396–418. Springer, 2024. doi: 10.1007/978-3-031-63498-7_24.

Chu Min Li and Felip Manyà. Maxsat, hard and soft constraints. In *Handbook of Satisfiability*, pp. 903–927. IOS Press, 2021. doi: 10.3233/FAIA201007.

Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, 1987. doi: 10.1016/0004-3702(87)90062-2.

Carsten Sinz. Towards an optimal cnf encoding of boolean cardinality constraints. In *Principles and Practice of Constraint Programming (CP 2005)*, volume 3709 of *Lecture Notes in Computer Science*, pp. 827–831. Springer, 2005. doi: 10.1007/11564751_73.

G. S. Tseitin. On the complexity of derivation in propositional calculus. In A. O. Slisenko (ed.), *Studies in Constructive Mathematics and Mathematical Logic, Part II*, pp. 115–125. 1968.

Leslie G. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979. doi: 10.1137/0208032.

Nathan Wetzler, Marijn J. H. Heule, and Warren A. Hunt Jr. Drat-trim: Efficient checking and trimming using expressive clausal proofs. In *Theory and Applications of Satisfiability Testing – SAT 2014*, volume 8561 of *Lecture Notes in Computer Science*, pp. 422–429. Springer, 2014. doi: 10.1007/978-3-319-09284-3_31.

A Proof Appendix Roadmap

This appendix provides full formal definitions and complete proofs for the main text claims. Per ICLR guidance, appendices may be arbitrarily long (reviewers are not required to read them). (Li & Manyà, 2021)

B Formal Preliminaries

B.1 Syntax and semantics (CNF)

A *literal* is a Boolean variable x or its negation $\neg x$. A *clause* is a disjunction of literals $C = \ell_1 \vee \cdots \vee \ell_r$. A *CNF formula* is a conjunction of clauses $F = \bigwedge_{t=1}^{m} C_t$.

An assignment $\mu : \mathrm{Var}(F) \to \{0, 1\}$ extends to literals by $\mu(\neg x) = 1 - \mu(x)$ and to clauses by disjunction. We write $\mu \models F$ if μ satisfies all clauses of F.

B.2 Weighted partial MaxSAT (WCNF)

A weighted partial MaxSAT instance consists of: (i) a set of *hard* CNF clauses H that must be satisfied, and (ii) a multiset of *soft* clauses $S = \{(C_u, \omega_u)\}$ with weights $\omega_u \geq 0$.

The objective is to find an assignment μ such that $\mu \models \bigwedge H$ and the satisfied soft weight $\sum_{(C_u,\omega_u)\in S} \omega_u \cdot \mathbf{1}[\mu \models C_u]$ is maximized. Equivalently, minimize total violated soft weight. See Li & Manyà (2021).

B.3 CNF normalization via Tseitin encodings

When φ is not in CNF, we use an equisatisfiable CNF encoding $\mathrm{CNF}(\varphi)$ (e.g. Tseitin-style) introducing auxiliary variables and linear overhead (Tseitin, 1968).

C Correctness of Gating and One-Hot Encodings

C.1 Clause-gating lemma

Lemma C.1 (Clause gating)**.** *Let z be a Boolean selector and let C be a clause. For any assignment μ, we have:*

$$\mu \models (\neg z \vee C) \quad \Longleftrightarrow \quad \mu(z) = 0 \text{ or } \mu \models C.$$

Proof. ($\Rightarrow$) If $\mu(z) = 1$ then $\neg z$ is false so C must be true. If $\mu(z) = 0$ we are done.

($\Leftarrow$) If $\mu(z) = 0$, then $\neg z$ is true, hence $(\neg z \vee C)$ is true. If $\mu(z) = 1$ and $\mu \models C$, then C is true and again $(\neg z \vee C)$ is true. □

Corollary C.2 (Gated CNF equivalence)**.** *Let* $\mathrm{CNF}(\varphi) = \bigwedge_{t=1}^{m} C_t$. *Then for any assignment* μ*,*

$$\mu \models \bigwedge_{t=1}^{m} (\neg z \vee C_t) \quad \Longleftrightarrow \quad \mu(z) = 0 \text{ or } \mu \models \mathrm{CNF}(\varphi).$$

Proof. Apply Lemma C.1 clausewise to all $t \in [m]$ and take conjunction. □

C.2 One-hot encoding correctness

For each question i, define selector variables $z_{i1}, \ldots, z_{iK}$.

We encode "exactly one" as:

$$\textsc{AtLeastOne}_i := (z_{i1} \vee \cdots \vee z_{iK}), \qquad \textsc{AtMostOne}_i := \bigwedge_{1 \le j < \ell \le K} (\neg z_{ij} \vee \neg z_{i\ell}).$$

Lemma C.3 (Exact-one correctness)**.** *An assignment* μ *satisfies* $\textsc{AtLeastOne}_i \wedge \textsc{AtMostOne}_i$ *iff there exists a unique* $j^\star(i) \in [K]$ *with* $\mu(z_{ij^\star(i)}) = 1$.

Proof. If μ satisfies $\textsc{AtLeastOne}_i$, at least one selector is true. If two distinct selectors were true, some pairwise clause in $\textsc{AtMostOne}_i$ would be violated. Hence exactly one is true. The converse direction is immediate. □

Remark C.4 (Linear-size one-hot)**.** The pairwise at-most-one costs $O(K^2)$ clauses. Sequential-counter encodings provide $O(K)$ size while preserving correctness (Sinz, 2005).

D Full Proof of MaxSAT Compilation Equivalence

D.1 Restating the coherent decoding problem

Given questions $q_1, \ldots, q_n$, candidates $(i, j) \in [n] \times [K]$, weights w_{ij}, semantic formulas φ_{ij}, and background theory B, define selection variables $s_{ij} \in \{0, 1\}$ with $\sum_j s_{ij} = 1$.

Define:

$$T(s) := B \cup \{\varphi_{ij} : s_{ij} = 1\}.$$

The goal is:

$$\max_s \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} s_{ij} \quad \text{s.t.} \quad \mathrm{SAT}(T(s)).$$

D.2 Weight shifting invariance

Lemma D.1 (Per-question shift invariance)**.** *Fix constants* $\alpha_i \in \mathbb{R}$. *Define* $w'_{ij} := w_{ij} + \alpha_i$. *Then for every feasible selection* s*,*

$$\sum_{i,j} w'_{ij} s_{ij} = \sum_{i,j} w_{ij} s_{ij} + \sum_{i=1}^{n} \alpha_i,$$

hence $\arg\max$ *over feasible* s *is unchanged.*

Proof. Using $\sum_j s_{ij} = 1$,

$$\sum_{i,j} w'_{ij} s_{ij} = \sum_{i,j} (w_{ij} + \alpha_i) s_{ij} = \sum_{i,j} w_{ij} s_{ij} + \sum_i \alpha_i \sum_j s_{ij} = \sum_{i,j} w_{ij} s_{ij} + \sum_i \alpha_i.$$

The last term is constant in s. □

D.3 WCNF CONSTRUCTION

We use selector variables z_{ij} (intended to equal s_{ij}). Hard clauses include:

- $\mathrm{CNF}(B)$;
- one-hot constraints for each i;
- for each (i, j) and each clause C in $\mathrm{CNF}(\varphi_{ij})$, the gated clause $(\neg z_{ij} \vee C)$.

Soft clauses include (z_{ij}) with nonnegative weights $\mathrm{Shift}(w_{ij})$ (via Lemma D.1).

Theorem D.2 (Compilation equivalence (full))**.** *Let $\mathcal{I}$ be the above WCNF instance. Then:*

1. *Any satisfying assignment μ of all hard clauses induces a feasible selection s^μ by $s^\mu_{ij} = \mu(z_{ij})$, such that $\mathrm{SAT}(T(s^\mu))$.*
2. *Among feasible selections, the MaxSAT objective equals the shifted preference score, hence any MaxSAT-optimal $\mu^\star$ yields an optimal coherent selection.*

Proof. (1) Fix μ satisfying all hard clauses. By Lemma C.3, for each i there is a unique $j^\star(i)$ with $\mu(z_{ij^\star(i)}) = 1$. Define $s^\mu_{ij} = \mu(z_{ij})$.

We show $\mathrm{SAT}(T(s^\mu))$ by exhibiting a model: Since $\mu \models \mathrm{CNF}(B)$, μ satisfies B (up to CNF equisatisfiability). For each chosen $(i, j^\star(i))$, $\mu(z_{ij^\star(i)}) = 1$; by Corollary C.2, μ must satisfy $\mathrm{CNF}(\varphi_{ij^\star(i)})$, hence $\varphi_{ij^\star(i)}$. Therefore μ satisfies all formulas in $B \cup \{\varphi_{ij} : s^\mu_{ij} = 1\}$, i.e. $T(s^\mu)$.

(2) The soft objective is

$$\sum_{i,j} \mathrm{Shift}(w_{ij}) \cdot \mathbf{1}[\mu \models z_{ij}].$$

Because z_{ij} is a unit clause, $\mathbf{1}[\mu \models z_{ij}] = \mu(z_{ij}) = s^\mu_{ij}$. Thus the objective equals $\sum_{i,j} \mathrm{Shift}(w_{ij}) s^\mu_{ij}$. By Lemma D.1, maximizing shifted weights is equivalent to maximizing original weights. Therefore any MaxSAT-optimal $\mu^\star$ yields an optimal coherent selection. □

E COMPLEXITY RESULTS (FULL PROOFS)

E.1 DECISION PROBLEM AND NP-COMPLETENESS

Define the decision version:

Definition E.1 (COHERENT-THRESHOLD)**.** Given $(B, \{\varphi_{ij}\}, \{w_{ij}\}, \tau)$ decide whether there exists a selection s such that $\mathrm{SAT}(T(s))$ and $\sum_{i,j} w_{ij} s_{ij} \geq \tau$.

Theorem E.2 (COHERENT-THRESHOLD is NP-complete)**.** *COHERENT-THRESHOLD is NP-complete, even when $K = 2$ and all weights are zero.*

Proof. Membership in NP: guess s, check $\sum_j s_{ij} = 1$ for all i, check $\mathrm{SAT}(T(s))$ via a SAT verifier, and compute the score in polynomial time.

NP-hardness: reduce from CNF-SAT. Given a CNF formula $F(x_1, \ldots, x_n)$, create n questions, one per variable. For question i, set $K = 2$ candidates with formulas:

$$\varphi_{i,1} := x_i, \qquad \varphi_{i,2} := \neg x_i.$$

Let $B := F$ and set all weights $w_{ij} = 0$ and $\tau = 0$. A selection s corresponds bijectively to a truth assignment to variables. Then $\mathrm{SAT}(T(s))$ holds iff the chosen literals satisfy F. Thus there exists a coherent selection iff F is satisfiable. □

E.2 OPTIMIZATION NP-HARDNESS VIA MAXSAT

Theorem E.3 (Optimization NP-hardness (explicit reduction))**.** *The optimization problem is NP-hard even when $K = 2$ and each φ_{ij} is either a single clause or $\top$.*

Proof. Reduce from weighted partial MaxSAT (Li & Manyà, 2021). Let H be the hard CNF and soft clauses $\{(C_t, \omega_t)\}_{t=1}^{m}$. Create m questions, each with two candidates:

$$\varphi_{t,1} := C_t \ \ (\text{weight } \omega_t), \qquad \varphi_{t,2} := \top \ \ (\text{weight } 0).$$

Set $B := H$. Any selection chooses a subset of soft clauses to include; feasibility requires satisfiability with H. Maximizing total weight matches the MaxSAT objective. □

F Belief-Revision Geometry (Full Proofs)

F.1 Selection space and distances

Let $\mathcal{S} := \{s \in \{0,1\}^{n \times K} : \sum_j s_{ij} = 1 \ \forall i\}$.

Let s^0 be a baseline selection (e.g. local top-1). Define a weighted edit distance:

$$d_\lambda(s, s^0) = \sum_{i=1}^{n} \sum_{j=1}^{K} \lambda_{ij} \cdot \mathbf{1}[s_{ij} \neq s^0_{ij}], \qquad \lambda_{ij} \geq 0.$$

Define coherent set $\mathcal{C} := \{s \in \mathcal{S} : \mathrm{SAT}(T(s))\}$.

Definition F.1 (Distance-based coherent revision operator)**.** Define:

$$\mathrm{Rev}_\lambda(s^0) := \arg\min_{s \in \mathcal{C}} d_\lambda(s, s^0),$$

breaking ties arbitrarily but deterministically.

Proposition F.2 (Core postulates)**.** *Assume $\mathcal{C} \neq \emptyset$. Then Rev_λ satisfies:*

1. ***Success:*** $\mathrm{Rev}_\lambda(s^0) \in \mathcal{C}$.
2. ***Vacuity:*** *if $s^0 \in \mathcal{C}$ then* $\mathrm{Rev}_\lambda(s^0) = s^0$.
3. ***Minimal change:*** *for any $s \in \mathcal{C}$,* $d_\lambda(\mathrm{Rev}_\lambda(s^0), s^0) \leq d_\lambda(s, s^0)$.

Proof. Success holds by construction. If $s^0 \in \mathcal{C}$, then $d_\lambda(s^0, s^0) = 0$ is minimal, proving vacuity. Minimal change is immediate from $\arg\min$ definition. □

F.2 Dalal-style specialization

Theorem F.3 (Dalal specialization (full))**.** *Suppose each question i corresponds to choosing a literal assignment for a dedicated world variable x_i: $K = 2$, $\varphi_{i,1} = x_i$, $\varphi_{i,2} = \neg x_i$. Let B be any CNF over $\{x_i\}$ and let $\lambda_{ij} = 1$. Then $\mathrm{Rev}_\lambda(s^0)$ returns a satisfying assignment of B with minimum Hamming distance to the baseline assignment, matching Dalal-style revision (Dalal, 1988).*

Proof. A selection $s \in \mathcal{S}$ encodes a unique truth assignment to all x_i. The constraint $s \in \mathcal{C}$ is exactly "assignment satisfies B". The edit distance $d_\lambda(s, s^0)$ equals the number of variables whose truth value differs from baseline, i.e. Hamming distance. Thus the minimizer is the nearest satisfying assignment, which is the Dalal construction. □

G MUS/MCS Duality and Repair Geometry

G.1 Grouped constraints

Let each candidate (i, j) correspond to a group of CNF clauses $G_{ij} := \mathrm{CNF}(\varphi_{ij})$, and let $G_B := \mathrm{CNF}(B)$. For a fixed selection s, define the group set:

$$\mathcal{G}(s) := G_B \ \cup \ \{G_{ij} : s_{ij} = 1\}.$$

A subset $\mathcal{U} \subseteq \mathcal{G}(s)$ is a *MUS (at group level)* if $\bigwedge \mathcal{U}$ is unsatisfiable and every strict subset is satisfiable. A subset $\mathcal{C} \subseteq \mathcal{G}(s) \setminus \{G_B\}$ is a *group-level correction set* if $\bigwedge(\mathcal{G}(s) \setminus \mathcal{C})$ is satisfiable.

G.2 HITTING SET DUALITY

Theorem G.1 (MCS as minimal hitting set of MUSes)**.** *Let $\mathcal{M}$ be the set of all group-level MUSes of $\mathcal{G}(s)$. A subset $\mathcal{C} \subseteq \mathcal{G}(s) \setminus \{G_B\}$ is a* minimal *correction set iff $\mathcal{C}$ is a* minimal hitting set *of $\mathcal{M}$ (i.e. it intersects every MUS, and no strict subset does). This is a standard diagnosis duality (Reiter, 1987).*

Proof. ($\Rightarrow$) Let $\mathcal{C}$ be a correction set. If $\mathcal{C}$ failed to intersect some MUS $\mathcal{U} \in \mathcal{M}$, then all groups in $\mathcal{U}$ would remain in $\mathcal{G}(s) \setminus \mathcal{C}$, making it unsatisfiable — contradiction. So $\mathcal{C}$ hits every MUS. Minimality follows similarly: if a strict subset $\mathcal{C}' \subset \mathcal{C}$ hit all MUSes, then removing only $\mathcal{C}'$ would restore satisfiability, contradicting minimality of $\mathcal{C}$.

($\Leftarrow$) Let $\mathcal{C}$ be a minimal hitting set of $\mathcal{M}$. Assume for contradiction that $\mathcal{G}(s) \setminus \mathcal{C}$ is unsatisfiable. Then it contains a MUS $\mathcal{U}$ (by finiteness and minimality), but $\mathcal{U} \subseteq \mathcal{G}(s) \setminus \mathcal{C}$ implies $\mathcal{U}$ is not hit by $\mathcal{C}$, contradiction. Thus removal of $\mathcal{C}$ restores satisfiability. Minimality follows from minimal hitting set property. □

G.3 FROM MUS/MCS TO ANSWER-LEVEL REPAIR

Theorem G.1 implies that repairing inconsistency is a hitting-set problem over MUSes. When grouped by answers, this yields an interpretable "which answers must change" view.

H COHERENCE GAP AS MINIMUM PREFERENCE SACRIFICE

Let $j_{\mathrm{loc}}(i) \in \arg\max_j w_{ij}$ and define $W_{\mathrm{loc}} = \sum_i w_{i\, j_{\mathrm{loc}}(i)}$. For any selection s, define the per-question sacrifice:

$$\Delta(s) := \sum_{i=1}^{n} \left(w_{i\, j_{\mathrm{loc}}(i)} - \sum_{j=1}^{K} w_{ij} s_{ij} \right).$$

Lemma H.1 (Coherence gap equals minimal sacrifice)**.**

$$\mathrm{CG} = \min_{s \in \mathcal{C}} \Delta(s).$$

Proof. By definition, $\mathrm{CG} = W_{\mathrm{loc}} - W^\star$ and $W^\star = \max_{s \in \mathcal{C}} \sum_{ij} w_{ij} s_{ij}$. Thus:

$$\mathrm{CG} = \min_{s \in \mathcal{C}} \left(W_{\mathrm{loc}} - \sum_{ij} w_{ij} s_{ij} \right) = \min_{s \in \mathcal{C}} \Delta(s).$$

□

I PROBABILISTIC EXTENSION: KL PROJECTION AND #P-HARDNESS

I.1 KL PROJECTION ONTO COHERENT SUPPORT

Define $Q_\theta(s) \propto \exp(\sum_{ij} w_{ij} s_{ij})$ on $\mathcal{S}$. Let $\mathcal{C} \subseteq \mathcal{S}$ be coherent set.

Theorem I.1 (KL projection is conditionalization)**.** *Assume $Q_\theta(\mathcal{C}) > 0$. Then*

$$P^\star = \operatorname*{arg\,min}_{P : \mathrm{supp}(P) \subseteq \mathcal{C}} \mathrm{KL}(P \| Q_\theta)$$

is $P^\star(s) = Q_\theta(s \mid s \in \mathcal{C})$.

Proof. For $s \in \mathcal{C}$, $Q_\theta(s) = Q_\theta(s \mid \mathcal{C}) Q_\theta(\mathcal{C})$. Thus

$$D_{\mathrm{KL}}(P \| Q_\theta) = \sum_{s \in \mathcal{C}} P(s) \log \frac{P(s)}{Q_\theta(s \mid \mathcal{C}) Q_\theta(\mathcal{C})} = D_{\mathrm{KL}}(P \| Q_\theta(\cdot \mid \mathcal{C})) - \log Q_\theta(\mathcal{C}).$$

The second term is constant in P, minimized when the first term is 0, i.e. $P = Q_\theta(\cdot \mid \mathcal{C})$. □

I.2 #P-HARDNESS OF COHERENT MASS COMPUTATION

Theorem I.2 (#P-hardness of $Q_\theta(\mathcal{C})$)**.** *Computing $Q_\theta(\mathcal{C})$ is #P-hard in general (Valiant, 1979).*

Proof. Reduce from #SAT. Given CNF $F(x_1, \ldots, x_n)$, create n questions with $K = 2$ candidates $\varphi_{i,1} = x_i$ and $\varphi_{i,2} = \neg x_i$. Let $B := F$ and set all weights $w_{ij} = 0$. Then Q_θ is uniform over all 2^n selections (assignments). The coherent set $\mathcal{C}$ corresponds exactly to satisfying assignments of F. Therefore

$$Q_\theta(\mathcal{C}) = \frac{\#\mathrm{SAT}(F)}{2^n}.$$

Computing $Q_\theta(\mathcal{C})$ thus computes $\#\mathrm{SAT}(F)$, which is #P-hard (Valiant, 1979). □

J CERTIFICATES: VERIFICATION GUARANTEES

J.1 FEASIBILITY CERTIFICATE

A feasibility certificate is a satisfying assignment μ to all variables in the compiled WCNF instance. Verification checks each hard clause evaluates to true under μ.

Proposition J.1 (Verification is linear-time in CNF size)**.** *Given CNF formula F and assignment μ, checking $\mu \models F$ takes $O(|F|)$ time.*

Proof. Evaluate each clause by scanning its literals; total work proportional to total literal occurrences. □

J.2 PROOF LOGGING CONTEXT

SAT proof logging and checking are standard (e.g. DRAT/DRAT-trim) (Heule, 2016; Wetzler et al., 2014). MaxSAT proof logging has become practical more recently, including certified preprocessing and equioptimality checking (Ihalainen et al., 2024).

K MULTI-ORACLE MAJORITY BOUND

Theorem K.1 (Majority vote error (Hoeffding bound))**.** *Let $O_1, \ldots, O_M$ be independent Bernoulli oracles with $\Pr[O_m \text{ correct}] = p > 1/2$. Let $\hat{Y}$ be majority vote. Then*

$$\Pr[\hat{Y} \neq Y] \leq \exp(-2M(p - 1/2)^2).$$

Proof. Let $X_m = \mathbf{1}[O_m \text{ correct}]$ so $\mathbb{E}[X_m] = p$. Majority error means $\sum_m X_m \leq M/2$. Apply Hoeffding's inequality to $\sum_m X_m$. □