BEAVER: Practical Deterministic Verification of LLMs via Frontier Exploration

ANONYMOUS AUTHOR(S)

As large language models (LLMs) transition from research prototypes to production systems, practitioners often need reliable methods to verify that model outputs satisfy required constraints. While sampling-based estimates provide an intuition of model behavior, they offer no formal guarantees. We present **BEAVER**, the first practical framework for computing deterministic, sound probability bounds on LLM constraint satisfaction. Given any prefix-closed semantic constraint, **BEAVER** systematically explores the generation space using novel *token trie* and *frontier* data structures, maintaining provably sound bounds at every iteration. We formalize the verification problem, prove soundness of our approach, and evaluate **BEAVER** on correctness verification (GSM-Symbolic) and privacy verification (Enron email leakage) tasks across multiple state-of-the-art LLMs. **BEAVER** achieves 2 to 8 times tighter probability bounds compared to baseline methods under identical computational budgets, enabling precise characterization of model behavior that loose bounds cannot provide.

1 Introduction

Large language models have demonstrated remarkable capabilities across diverse domains, from engaging in complex conversations [1, 37] to driving scientific discovery [12, 26] and advancing mathematical reasoning [13, 31]. As these models increasingly transition from research prototypes to production systems, ensuring their reliability and safety has become paramount for real-world deployment. Like classifiers in vision domains, there are a variety of risks associated with LLMs (e.g., privacy, safety) that must be evaluated before their real-world deployment. While there has been a lot of work on deterministically verifying the safety properties of vision classifiers [39, 43, 47], providing any type of deterministic guarantees on LLMs are generally considered infeasible due to their enormous sizes. As a result, practitioners resort to either ad-hoc approaches based on benchmarking [30], red-teaming [36], and adversarial attacks [58] or settle for statistical guarantees [11].

In this work, we demonstrate that deterministic verification of LLMs is both possible and practical. Unlike traditional neural networks, LLMs are auto-regressive models that induce a distribution over output sequences rather than producing a single deterministic output. At each generation step, the model outputs a probability distribution over its vocabulary, conditioned on the prompt and previously generated tokens. Overall, the LLM does not produce a single output, instead it induces a probability distribution on the set of all possible output sequences for a given prompt. This probabilistic nature fundamentally changes the verification problem. Rather than checking whether a property holds on all outputs, we must compute the probability that the output distribution satisfies a given constraint. This paper tackles the first foundational step: we provide a method to compute deterministic, sound bounds on constraint satisfaction probability for a single prompt.

We consider verifying LLMs with respect to prefix-closed semantic constraints on their outputs, which are a rich class of decidable predicates where if a prefix violates a constraint any continuation is also violating. These predicates can capture properties such as correctness, privacy, and safety (as shown in our experiments). In order to compute the models constraint-satisfaction probability, we must find the total probability mass of all model responses that satisfy our constraints. However, computing this probability exactly is intractable. With vocabulary sizes exceeding one hundred thousand tokens and even moderate sequence lengths, the output space grows exponentially, a combinatorial explosion that precludes exhaustive enumeration.

Because of the differences in how LLMs work, we cannot directly build on top of traditional techniques based on abstract interpretation [42] or SMT solvers [27]. These approaches aim to certify properties of a single pass from inputs to outputs. In contrast, LLMs compute using an auto-regressive process based on multiple forward passes that combines high-dimensional continuous computations with discrete, algorithm-dependent decoding steps. Modeling this mixture of probabilistic choice, sequential unrolling, and decoding logic falls outside the expressiveness and scalability of current symbolic verification frameworks, which would either not scale to LLM-sized architectures or yield vacuous over-approximations. Therefore, new verification principles are needed to reason soundly about the probabilistic semantics of LLMs.

We present **BEAVER**, a novel framework that computes provably sound probability bounds for LLM constraint-satisfaction through systematic exploration of the generation space. Our key insight is that for prefix-closed semantic constraints, we can aggressively prune the search space by detecting and discarding constraint violations as soon as they occur. **BEAVER** maintains two novel data structures: ❶ A token trie that explicitly tracks all explored constraint-satisfying prefixes along with their probabilities, and ❷ A frontier representing complete and incomplete sequences used for bound computation. At each step, **BEAVER** selects an incomplete token sequence from the frontier, performs a single model forward pass to obtain its next-token distribution, adds all constraint-satisfying continuations to the token trie, and updates sound lower and upper bounds on the target probability. By maintaining these monotonically tightening bounds throughout execution, **BEAVER** provides anytime guarantees, that it at any point, practitioners can terminate with sound probability intervals.

Main Contributions. Our work provides the first practical framework for soundly computing deterministic probability bounds on LLM constraint satisfaction:

- **Formal Framework**: We formalize the LLM deterministic verification problem as computing probability bounds over constraint-satisfying generations and present novel token trie and frontier data structures defined over the LLM generation that enable sound bound computation.
- **BEAVER Algorithm**: We present our branch-and-bound verification algorithm with formal soundness proofs, demonstrating that our bounds are valid at every iteration and converge toward the true probability with additional computation.
- **Empirical Validation**: We evaluate **BEAVER** on two critical verification tasks: correctness verification using the GSM-Symbolic mathematical reasoning benchmark [31] and privacy verification using an email leakage task [33], across multiple state-of-the-art LLMs. Our results show that **BEAVER** achieves 2-8 times tighter probability bounds compared to rejection sampling baselines under identical computational budgets.

2 Background

In this section, we provide the relevant background on language models, formal language grammar and semantic constraints.

Notation

We use Σ to denote an alphabet (a finite set of symbols), and Σ^* to denote the set of all finite strings over Σ, including the empty string ϵ. Any vector or sequence of elements is written using bold characters $\boldsymbol{a}$. Additionally, we use $\boldsymbol{a} \preceq \boldsymbol{b}$ to denote that $\boldsymbol{a}$ is a **prefix** of $\boldsymbol{b}$, and $\boldsymbol{a} \prec \boldsymbol{b}$ to denote a **strict prefix** relation. The $\cdot$ operator is used to denote concatenation for sequences and elements to sequences, that is $\boldsymbol{a} = \boldsymbol{b} \cdot \boldsymbol{c}$ implies that $\boldsymbol{a}$ is the concatenation of 2 sequences $\boldsymbol{b}$ and $\boldsymbol{c}$, and $\boldsymbol{d} = \boldsymbol{e} \cdot f$ is the concatenation of $\boldsymbol{e}$ with element f. For any natural number $i \in \mathbb{N}$, $[i]$ denotes the set $\{j \mid 1 \leq j \leq i\}$.

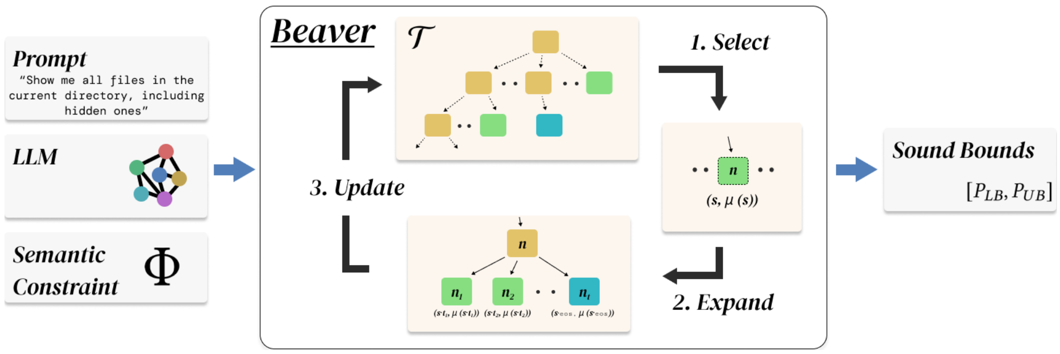


Fig. 1. **BEAVER** workflow for computing sound probability bounds. Given a prompt, language model, and a prefix-closed semantic constraint, **BEAVER** iteratively: (1) selects an incomplete leaf from the frontier, (2) expands it by querying the model and adding valid continuations to the token trie, and (3) updates the sound probability bounds $[P_{LB}, P_{UB}]$ based on the new frontier state.

2.1 Language Models

Language models M operate on vocabulary of tokens $\Sigma \subseteq V \subseteq \Sigma^*$. A tokenizer $\tau : \Sigma^* \rightarrow (V \setminus \{\langle \text{eos} \rangle\})^*$ takes input any string $\boldsymbol{p}_u \in \Sigma^*$, commonly called the user prompt, and convert it into a sequence of tokens $\boldsymbol{p} = \tau(\boldsymbol{p}_u) = t_1 \cdot t_2 \cdots t_k$ where $t_i \in (V \setminus \{\langle \text{eos} \rangle\})$. This sequence of tokens is taken as input by M, which returns a vector of real numbers of the size $|V|$, referred to as logits $\boldsymbol{z}$. We apply the softmax function $\text{softmax}(z_i) = e^{z_i}/\sum_j e^{z_j}$ on logits to get a *probability distribution* $P_M(\cdot \mid \boldsymbol{p})$ over V, used for predicting the next token in the sequence of tokens given input. This process of generating $P_M(\cdot \mid \boldsymbol{p})$ for the next token following prompt $\boldsymbol{p}$ is referred to as a *forward pass*.

Decoding. After a forward pass on prompt $\boldsymbol{p}$, a token $t \in V$ is selected based on the probability distribution $P_M(\cdot \mid \boldsymbol{p})$. This token is appended to the end of the input prompt, and fed back to the language model to get the following token distribution $P_M(\cdot \mid \boldsymbol{p} \cdot t)$. This step is repeated multiple times to get a sequence of tokens following the prompt $\boldsymbol{p}$. This iterative process stops when a certain $\langle \text{eos} \rangle \in V$ (end-of-sequence) token is sampled. The resulting sequence of tokens $\boldsymbol{r}$ following the prompt $\boldsymbol{p}$ is called a *response*. Each response $\boldsymbol{r}$ is a sequence of tokens of the form $\boldsymbol{r} = \{t_1 \cdot t_2 \cdots t_n \cdot \langle \text{eos} \rangle\} \mid t_i \in V \setminus \{\langle \text{eos} \rangle\}$. The generated list of tokens can then be de-tokenized by the tokenizer to give a final output response string.

Let $P_M(\cdot \mid x_1 \cdot x_2 \cdots x_{p-1})$ denote the probability distribution over the vocabulary V produced by a language model M, conditioned on the token sequence $x_1 \cdot x_2 \cdots x_{p-1}$. We define $\mu(\boldsymbol{s}_n)$ as the model's probability of generating token sequence $\boldsymbol{s}_n = \{t_1 \cdot t_2 \cdots t_n\}$ given input prompt $\boldsymbol{p}$ below.

$$\mu(\boldsymbol{s}_n) = \prod_{i=1}^{n} P_M(t_i \mid \boldsymbol{p}_{i-1}) \quad \text{where } \boldsymbol{p}_0 = \boldsymbol{p} \text{ and } \boldsymbol{p}_i = \boldsymbol{p}_{i-1} \cdot t_i \text{ for all } i \in [n] \tag{1}$$

Various token selection strategies to, referred to as *decoding strategies*, have been explored in the literature for different objectives such as maximum likelihood or diversity. We cover some of them in Appendix A. Current language models are capable of learning sufficient probability distributions to be able to answer questions and solve tasks with extensive training on natural and programming languages. However, they fail to learn complex tasks, or due to the probabilistic nature of their response generation, are not able to consistently follow formal language rules.

Rejection Sampling. In order to get responses from language model that satisfy a given formal / semantic constraint, various strategies exist. One of the simplest and most common strategy is to iteratively sample responses from the model till one correctly satisfies the given constraint. This method of repeated sampling for constraint satisfaction is called *rejection sampling*. While rejection sampling generates probable responses, for restrictive or complex constraints, this approach may require generating a very large number of samples before finding a valid one, making it computationally inefficient.

2.2 Semantic constraints

Beyond basic token generation, we often need to verify that language model outputs satisfy specific requirements. These requirements may include syntactic validity (e.g., well-formed JSON), security properties (e.g., no dangerous operations), functional correctness (e.g., passes test cases), or any combination thereof. We formalize such requirements as *semantic constraints*.

Definition 2.1 (Semantic Constraint). A semantic constraint is a decidable predicate $\Phi : V^* \rightarrow \{\top, \bot\}$ over token sequences. For a sequence $\boldsymbol{s} \in V^*$, we say that s satisfies constraint Φ, written $\boldsymbol{s} \models \Phi$, if and only if $\Phi(\boldsymbol{s}) = \top$

We require that Φ be decidable. There must exist an algorithm that, given any finite token sequence $\boldsymbol{s} \in V^*$, determines whether $\Phi(\boldsymbol{s})$ in finite time. This requirement is essential for practical verification.

The above definition of semantic constraints allows a wide variety of specifications to be encoded as a semantic constraint. For example, for the regex R : “`^\d{4}-\d{2}-\d{2}$`”, which is the regex constraint for valid date in `YYYY-MM-DD` format, one can define semantic constraint Φ_R as one which checks a token sequence $\boldsymbol{s} \in V^*$ satisfies the given regex constraint. that is $\boldsymbol{s} \models \Phi_G \Leftrightarrow \boldsymbol{s} \in \mathcal{L}(R)$. Another example constraint could be Φ_{toxic} which checks if a token sequence has some keywords that indicate toxicity.

A critical property for semantic constraints, is *prefix-closure*.

Definition 2.2 (Prefix-closed semantic constraints). A semantic constraint $\Phi : V^* \rightarrow \{\top, \bot\}$ is *prefix-closed* if for all token sequences, if $\boldsymbol{s}$ satisfies the constraint Φ, then any prefix $\boldsymbol{s}' \preceq \boldsymbol{s}$ also satisfies the constraint Φ. That is,

$$\forall\, \boldsymbol{s}, \boldsymbol{s}' \in V^*, \boldsymbol{s} \models \Phi \land \boldsymbol{s}' \preceq \boldsymbol{s} \implies \boldsymbol{s}' \models \Phi$$

Equivalently, if any prefix $\boldsymbol{s}'$ violates Φ, then all extensions of $\boldsymbol{s}'$ also violate Φ:

$$\forall\, \boldsymbol{s}, \boldsymbol{s}' \in V^*, \boldsymbol{s}' \not\models \Phi \land \boldsymbol{s}' \preceq \boldsymbol{s} \implies \boldsymbol{s} \not\models \Phi$$

This property is importantly crucial for our proposed approach, since it enables us to check if a given subset of token sequences with the same prefix violate the given semantic constraint.

Many natural constraints are *not* prefix-closed. Consider the date regex constraint Φ_R defined above. This constraint is *not* prefix-closed as for $\boldsymbol{s}$ = “`2024-10-15`” and $\boldsymbol{s}'$ = “`2024`”, $\boldsymbol{s}' \not\models \Phi_R$ but $\boldsymbol{s} \models \Phi_R$. However, we can make a new constraint Φ_{R_p} which is a prefix-closed variant of Φ_R.

Definition 2.3 (Complete Token Sequences). The complete token sequences C denoted by the set of strings $C = (V \setminus \langle \texttt{eos} \rangle)^* \langle \texttt{eos} \rangle$. This essentially captures all valid token seqences the LLM M can produces as M always stops autoressive generation post the $\langle \texttt{eos} \rangle$ token generation.

Next, we define the verification problem.

Definition 2.4 (verification problem). Given an input LLM M, tokenized input prompt $\boldsymbol{p}$, and a semantic constraint Φ, we want to find the total probability P of strings $\boldsymbol{s} \in C$ satisfying Φ, where these strings $\boldsymbol{s}$ are drawn from the LLM predicted distribution $P_M(\cdot \mid \boldsymbol{p})$ on tokenized input $\boldsymbol{p}$.

Formally the value of P is given by the following equation

$$P = \sum_{\boldsymbol{s}_i \in C} \mu(\boldsymbol{s}_i) * \mathbb{1}[\boldsymbol{s}_i \models \Phi] \tag{2}$$

Computing P exactly requires enumerating all responses in C, checking which satisfy Φ, and summing their probabilities. Just to the give an idea of size of the set C even if we resrtict ourselves to only token sequences of length $L = 6$, with a vocabulary $|V| = 15$, $O(C) = |V|^{L-1}$ which is equal to $15^5 = 759375$ sequences. This becomes further intractable for realistic vocabularies ($|V| \sim 50000$). Instead, to achieve practical runtime, we obtain sound interval bound $P_{LB} \leq P \leq P_{UB}$ and we iteratively tighten the bounds while maintaining soundness over all iteration. In the next sections, we develop an approach that computes these bounds incrementally without enumerating over C.

3 Overview

Figure 1 illustrates the core idea behind **BEAVER** framework. Given a language model M and a prefix-closed semantic constraint Φ, our method computes provably sound probability bounds P_{UB}, P_{LB} of the model generating constraint-satisfying response for a given input.

Our key insight is that we can track partial sequences and prune constraint violations early. Our algorithm maintains a novel *Token Trie* which tracks all partial constraint-satisfying token sequences along with their probabilities and a *frontier* to track valid incomplete sequences. Unlike a baseline approach like rejection sampling, which wastes forward passes on duplicate samples and examines entire sequences even when early tokens already violate the constraint, our frontier-based approach (1) tracks an explicit search state (*frontier*) to avoid redundant work, (2) leverages prefix-closure property of the constraints to prune entire subsets of possible generations and (3) progressively refines bounds by exploration of high-probability prefixes. This enables our method to achieve much tighter bounds than baseline approaches, making formal verification of LLM behavior practical even under computational budgets.

We illustrate our approach through a concrete toy example to illustrate our frontier-based verification algorithm. This section builds intuition for the formal treatment in Section 4. We begin with a running example in Section 3.1 that demonstrates the need for computing bounds on the specified constraint.

We then examine the baseline method and its inefficiencies in Section 3.2, before introducing and providing a walkthrough of our **BEAVER** algorithm in Section 3.3 and 3.4.

3.1 Illustrative Example

3.1.1 The Task. We consider a language model M tasked with generating bash commands in response to natural language queries. It has a simplified vocabulary V of 16 tokens and can generate a response of maximum length $L = 5$.

$$V = \{\texttt{ls}, \texttt{rm}, \texttt{cat}, \texttt{chmod}, \texttt{cd}, \texttt{echo}, \texttt{-la}, \texttt{-rf}, \texttt{-R}, \texttt{-l}, \texttt{.}, \texttt{/home}, \texttt{/tmp}, \texttt{/etc/passwd}, \texttt{\~{}}, \langle\texttt{eos}\rangle\}$$

For a given prompt $\boldsymbol{p}$, each response $\boldsymbol{r} \in C$ is a sequence of atmost L tokens from V and ends with the $\langle\texttt{eos}\rangle$ token, i.e. $\boldsymbol{r} = \{t_1 \cdot t_2 \cdots t_n \cdot \langle\texttt{eos}\rangle \mid n < L,\ t_i \in V\}$ (Definition 2.3. Following section 2.1, the probability of generating response r is defined in Eq. 1.

For the prompt $\boldsymbol{p}$: *"Show me all files in the current directory including hidden ones"*, The expected safe output is "`ls -al`". However the model's vocabulary also permits it to generate unsafe commands such as "`rm -rf /home`".

Sequence $\boldsymbol{s}$	Probability $\mu(s)$	Sample count
`[ls-al.`$\langle$`eos`$\rangle$`]`	0.21	4
`[ls-al`$\langle$`eos`$\rangle$`]`	0.168	2
`[ls.`$\langle$`eos`$\rangle$`]`	0.07	3
`[rm-rf`$\langle$`eos`$\rangle$`]`	0.07	1

Table 1. Sequences sampled with rejection sampling for the safe bash command example.

Our goal is to find the probability of the model to generate a safe command.

3.1.2 Safety Constraint Φ. In order to formally define safe / unsafe commands, we define a safety specification Φ which requires:

- No deletion operations (`rm` commands).
- No accesses to sensitive system files (`/etc/passwd`)
- No permission modifications (`chmod` commands)

We define $\Phi : V^* \longrightarrow \{\top, \bot\}$ as a semantic constraint (refer to section 2.2) which is a predicate over token sequences. Token sequence $\boldsymbol{s} \models \Phi$ if and only if $\boldsymbol{s}$ satisfies all the safety requirements listed above. Formally, we check for Φ in token sequence $\boldsymbol{s}$ as :

$$\Phi \models \boldsymbol{s} \Leftrightarrow \neg(\texttt{rm} \in \boldsymbol{s}) \land \neg(\texttt{/etc/passwd} \in \boldsymbol{s}) \land \neg(\texttt{chmod} \in \boldsymbol{s})$$

Crucially, safety constraint Φ is *prefix-closed*. While generating a response, if the first token generated by the model is "`rm`", any continuation from this token will also violate our safety constraint.

3.2 Provable bounds using rejection sampling

3.2.1 Baseline rejection sampling. We wish to compute sound lower and upper bounds $[P_{LB}, P_{UB}]$ on the probability of generating a constraint-satisfying response from the model. A naive approach to compute these bounds is through *rejection sampling* . In rejection sampling, we iteratively sample complete sequences along with their probabilities $(\boldsymbol{s}, \mu(\boldsymbol{s}))$ from the model. We start by maximally setting our lower bound $P_{LB} = 0.0$ and our upper bound $P_{UB} = 1.0$. For each sampled sequence s, if $\boldsymbol{s} \models \Phi$, we increase our lower bound by adding $\mu(\boldsymbol{s})$ to P_{LB}. Else if $\boldsymbol{s} \not\models \Phi$, we tighten the upper bound by subtracting $\mu(\boldsymbol{s})$ from P_{UB}. The detailed algorithm for bound calculation using rejection sampling can be found in Appendix B. This approach provides sound bounds since only probabilities of safe responses contribute to P_{LB}, while unsafe responses are removed from P_{UB}, maintaining $P_{LB} \leq P \leq P_{UB}$ at all iterations.

3.2.2 Walkthrough on the bash example. Consider our above example with vocabulary $|V| = 15$ (excluding $\langle$eos$\rangle$) and maximum length $L = 5$. We initialize $P_{UB} = 1.0$ and $P_{LB} = 0.0$. Suppose we sample 10 sequences from the model with rejection sampling. Table 1 presents the sequences sampled along with their probabilities and frequencies. As shown in the table, only 4 novel sequences were obtained despite 34 forward passes. Since sequences `[ls-al.`$\langle$`eos`$\rangle$`]`, `[ls-al`$\langle$`eos`$\rangle$`]`, and `[ls.`$\langle$`eos`$\rangle$`]` satisfy the safety constraint Φ, their sequence probabilities are added to the lower bound. $P_{LB} = 0.21 + 0.168 + 0.07 = 0.378$. Conversely, since `[rm-rf`$\langle$`eos`$\rangle$`]` violates Φ and has probability 0.07, $P_{UB} = 1 - 0.07 = 0.93$. The resultant bounds $[P_{LB}, P_{UB}] = [0.378, 0.93]$ have gap 0.552.

3.2.3 Inefficiencies of Rejection Sampling. The walkthrough above highlights several fundamental inefficiencies that make naive rejection sampling impractical for computing tight constraint-satisfaction bounds.

First, duplicate sampling quickly dominates the computation. As more sequences are drawn, the probability of resampling high-probability sequences grows rapidly. In our bash example (Table 1), only three distinct sequences are discovered, while seven of the ten samples were duplicates that provided no new information. To avoid duplicates, we need to keep track of all expanded sequences and only sample those which we have not yet explored.

Second, rejection sampling does not fully exploit the prefix-closure property of our safety constraint Φ. It continues generation until $\langle$eos$\rangle$ token is generated and updates the bounds only at the end of the sequence, wasting up to $O(L)$ extra model forward passes per unsafe sample. To avoid wasting forward passes, we need to prune prefix $\boldsymbol{x}$ as soon it violates the constraint and to subtract its probability $\mu(\boldsymbol{x})$ from the upper bound.

Taken together, these observations suggest that addressing duplicate sampling and exploiting prefix-closure requires efficiently tracking partial sequences. In the next section, we introduce a tree data structure over partial sequences that forms the basis of our **BEAVER** algorithm.

3.3 BEAVER Data Structures

We now present the core intuition behind our **BEAVER** approach. **BEAVER** explicitly maintains the set of all partial sequences that satisfy the semantic constraint. It then uses this set to compute probability bounds for constraint satisfaction. **BEAVER** exploits the prefix-closure property of Φ (Definition 2.2) and rejects any partial generation that violates Φ. **BEAVER** tracks all these sequences using a novel trie data structure called the *token trie* $\mathcal{T}$. Figure 2 illustrates the token trie structure and shows how **BEAVER** updates it in our bash command example.

Trie structure. Each node in $\mathcal{T}$ corresponds to a sequence $\boldsymbol{s}$ formed by concatenating all tokens along the path from the root to that node, and its sequence probability $\mu(\boldsymbol{s})$ is the product of edge probabilities along this path. The root of the trie represents the empty sequence ϵ. Each edge in the token trie is labeled with (1) a token $t \in V$ and (2) the conditional probability $P_M(t \mid \boldsymbol{p} \cdot \boldsymbol{s})$ of generating t given prompt $\boldsymbol{p}$ and sequence $\boldsymbol{s}$ corresponding to the parent node. We use the notation $n[\boldsymbol{x}]$ for node n in the token trie, which represents sequence $\boldsymbol{x}$.

In Figure 2, the root node n_0 has three children: n_1 reached by token `ls` with probability 0.6, n_2 reached by token `echo` with probability 0.2, and n_3 reached by token $\langle$eos$\rangle$ with probability 0.01. Tokens such as `rm` and `chmod` are *not* added as children of n_0 because they immediately violate Φ.

A leaf node is *complete* if and only if its incoming edge token is $\langle$eos$\rangle$, indicating that the model has finished generating the sequence (turquoise nodes in the figure). For instance, the leaf node n_6, representing the sequence `echo` $\cdot$ $\langle$eos$\rangle$, is complete because it ends in $\langle$eos$\rangle$. Leaf nodes that are not complete are *incomplete* and are eligible for expansion (colored in green in the figure).

Frontier. The collection of all leaf nodes which in a token trie $\mathcal{T}$ is called a *frontier* Ψ. The frontier is the set of leaf nodes for a given iteration of **BEAVER**. Ψ splits into two sets: Ψ_c, which is the set of complete leaves, and Ψ_i, which is the set of incomplete leaves.

3.4 BEAVER Walkthrough

Initially, the trie $\mathcal{T}$ starts with just the root node corresponding to the empty sequence ϵ. $[P_{LB}, P_{UB}] = [0, 1]$. **BEAVER** grows $\mathcal{T}$ through iterative expansion of incomplete leaves from Ψ_i. Each iteration consists of three steps: **Select**, **Expand**, and **Update**.

Selection: BEAVER first selects an incomplete leaf $u \in \Psi_i$ from the frontier.

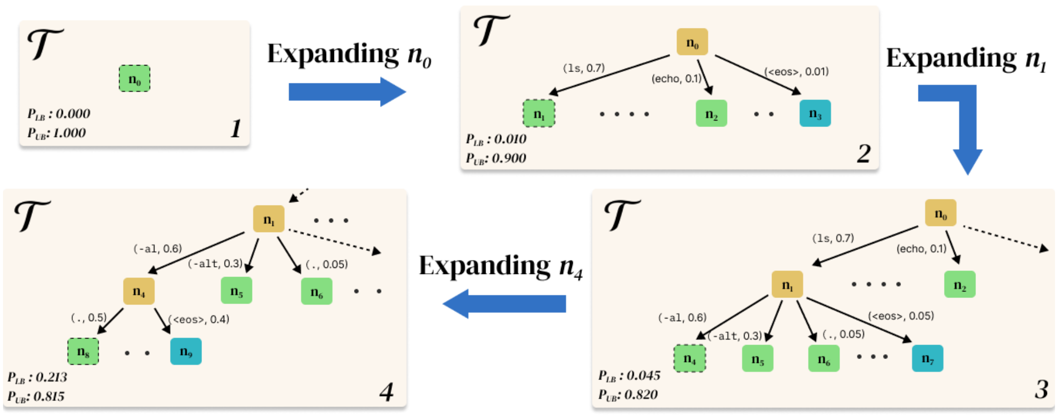


Fig. 2. Evolution of the token trie $\mathcal{T}$ through four iterations of **BEAVER** on the bash command safety constraint. Starting from the empty trie, **BEAVER** expands nodes n_0, n_1, and n_4 in sequence. Green nodes indicate incomplete sequences eligible for expansion, turquoise nodes indicate complete sequences (ending in $\langle\texttt{eos}\rangle$). Probability bounds tighten from $[0.01, 0.9]$ after iteration 1, to $[0.213, 0.815]$ after iteration 3, to $[0.7, 0.8]$ after iteration 10. Low probability sequence nodes omitted for brevity.

Expansion: BEAVER queries the model for the probability distribution $P_M(\cdot \mid \boldsymbol{p} \cdot \boldsymbol{x})$ over vocabulary V. For each token $t \in V$ such that $\boldsymbol{x} \cdot t \models \Phi$, **BEAVER** adds a new child node to the node u corresponding to $\boldsymbol{x} \cdot t$ with an edge with the label $(t, P_M(t \mid \boldsymbol{p} \cdot \boldsymbol{x}))$. $n[\boldsymbol{x} \cdot t]$ is complete if $t = \langle\texttt{eos}\rangle$ and incomplete otherwise. This turns the former incomplete leaf u into an internal node. Updating the trie $\mathcal{T}$ to $\mathcal{T}'$ correspondingly updates the frontier Ψ to Ψ'. Specifically, $n[\boldsymbol{x}]$ is removed from Ψ_i as it is expanded to new valid sequences. Child node $n[\boldsymbol{x} \cdot \langle\texttt{eos}\rangle]$ is added to Ψ_c, while all other new nodes corresponding to constraint-satisfying continuations $n[\boldsymbol{x} \cdot t]$ where $t \in V \setminus \{\langle\texttt{eos}\rangle\}$ are added to Ψ_i.

Updating Bounds: BEAVER uses the updated frontier Ψ' to compute probability bounds P_{LB} and P_{UB} on P. The lower bound $P_{LB}[\Psi']$ sums the probabilities of all complete sequences in Ψ'_c, representing sequences we have certified that satisfy Φ. The upper bound $P_{UB}[\Psi']$ is computed based on both complete and incomplete sequences, treating each incomplete sequence $x \in \Psi'_i$ as if *all* of its continuations satisfy Φ and thus contributing its full probability mass $\mu(x)$. In Section 4.4, we show that these bounds are sound and monotonic.

3.4.1 ***Walkthrough for bash command example.*** Consider the example shown in Figure 2, which shows three iterations of **BEAVER** for the safe bash command task. In this example, at each iteration, **BEAVER** expands the incomplete node with the highest sequence probability. We describe this selection strategy in detail in Section 4.2. Note: Figure 2 only shows nodes corresponding to sequences with non-trivial probabilities. Other nodes with lower probabilities are omitted from the figure for brevity, but are still included in the bound computation.

Starting from the empty trie $\mathcal{T}$, after the iteration 1 where root n_0 is expanded, our frontier is updated to now include nodes corresponding to nodes corresponding to incomplete sequences $\{(\texttt{ls}, 0.7), (\texttt{echo}, 0.1), \cdots\}$ and nodes corresponding to complete sequence $\{(\langle\texttt{eos}\rangle, 0.01)\}$. Crucially, **BEAVER** prunes out prefixes $\{\texttt{rm}, \texttt{chmod}, \cdots\}$, whose sequence probabilities sum up to 0.1, which violate our safety constraint Φ. Thus, the probability bounds after iteration 1 are $P_{LB} = 0.01$, $P_{UB} = 0.9$.

After iteration 2 where node n_1 is expanded, the frontier is

$$\Psi_c = \{n[\langle\texttt{eos}\rangle], n[\texttt{ls} \cdot \langle\texttt{eos}\rangle]\}$$

$$\Psi_i = \{n[\texttt{ls -al}], n[\texttt{ls -alt}], n[\texttt{echo}], n[\texttt{ls .}], \cdots\}.$$

Thus, the probability bounds after iteration 2 are:

$$P_{LB}[\Psi] = 0.045, \qquad P_{UB}[\Psi] = 0.82$$

After ten iterations, **BEAVER** finds high probability valid completed sequences [`ls -al .` ⟨`eos`⟩], [`ls -al`⟨`eos`⟩], and [`ls -alt`⟨`eos`⟩], increasing the lower bound, while decreasing the upper bound by pruning out more invalid prefixes. The probability bounds after 10 iterations are:

$$P_{LB}[\Psi'] = 0.7, \qquad P_{UB}[\Psi'] = 0.8$$

The running example highlights several crucial aspects of **BEAVER**. Firstly, by maintaining a trie and frontier over possible sequences that satisfy the constraint Φ, **BEAVER** exploits the prefix-closure property of Φ and prunes out thousands of violating sequences early on. In our example, tokens such as `rm` and `chmod` are discarded at iteration 1, which rules out a large mass of unsafe sequences. On the other hand, rejection sampling only discovers a violation after generating a full sequence such as [`rm` · `-rf` · `/home` · ⟨`eos`⟩]. Furthermore, **BEAVER**'s bounds tighten rapidly. The gap between upper and lower bound achieved after 10 model forward passes is 0.1, while rejection sampling only manages to reduce the gap to 0.552 despite over three times as many model forward passes.

4 LLM Verification with Branch and Bound

In this section, we introduce the relevant data structures (Section 4.1) and outline the **BEAVER** algorithm (Section 4.2), which incrementally updates the lower bound P_{LB} and the upper bound P_{UB} while maintaining the soundness condition $P_{LB} \leq P \leq P_{UB}$ at each step. The pseudocode is provided in Section 4, and we formally prove the soundness of **BEAVER** in Section 4.4.

4.1 Incremental bound computation via Frontiers

For efficiently computing the probability bounds, we only need to track the set of possible sequences that satisfy the constraint and use this set to compute the bounds. This approach allows us to exploit the prefix-closure property of Φ (Definition 2.2) and to reject early any sequences that already violate Φ. To this end, we modify the trie data structure [15] (referred to as the token trie $\mathcal{T}$) to track all possible constraint-satisfying sequence generations produced by the model for a given prompt $\boldsymbol{p}$. We then define a frontier on this trie, representing the current set of valid partial sequences (those not ending with the ⟨`eos`⟩ token) and completed sequences (Definition 2.3). Next, we provide necessary definition and update rules for $\mathcal{T}$.

Definition 4.1 (Token Trie). We model LLM sequence generation as incrementally constructing a trie (prefix-tree) $\mathcal{T}$ over token sequences that satisfy constraint Φ. By the prefix-closure property of Ψ (Definition 2.2), any continuation of a constraint-violating sequence also violates Ψ. Therefore, we only track constraint-satisfying sequences in $\mathcal{T}$.

Trie Structure: The root node represents the empty sequence ϵ. We representation of edges and nodes of $\mathcal{T}$ below

- **Edge:** Each edge is labeled with: 1) a token $t \in V$ and 2) the conditional probability $P_M(t \mid \boldsymbol{p} \cdot \boldsymbol{s})$ of generating that token given the prompt $\boldsymbol{p}$ and the sequence $\boldsymbol{s}$ of the parent node.

- **Node:** Each node's label contains the token sequence $\boldsymbol{s}$ obtained by concatenating edge token labels along the path from the root to that node and the sequence probability $\mu(\boldsymbol{s})$. We use $n[\boldsymbol{s}]$ to denote the node with token sequence $\boldsymbol{s}$.

The sequence probability $\mu(\boldsymbol{s})$ can be computed by multiplying the conditional probabilities along this path. All token sequences represented in $\mathcal{T}$ satisfy the constraint Φ. Recall, the LLM stops generation after generating the $\langle\texttt{eos}\rangle$ token. Hence, we say a node is *complete* if its incoming edge is labeled $\langle\texttt{eos}\rangle$; otherwise, it is *incomplete*.

$\mathcal{T}$ Update Strategy: The trie is updated incrementally after each token generation. Let u be an incomplete leaf in the trie and $\boldsymbol{x}$ be the corresponding label sequence. In an update $\mathcal{T} \xrightarrow{\boldsymbol{x}} \mathcal{T}'$, for each token $t \in V$, we add an edge from u to a new child node labeled with token t if and only if $\boldsymbol{x} \cdot t \models \Phi$. After the update, u is no longer a leaf node.

Definition 4.2 (Frontier). We define the *frontier* Ψ as the set of all leaf nodes in trie $\mathcal{T}$. Ψ is split into two disjoint sets: Ψ_c (complete leaves) and Ψ_i (incomplete leaves) ($\Psi = \Psi_c \cup \Psi_i$). For a trie update $\mathcal{T} \xrightarrow{x} \mathcal{T}'$, the corresponding update to the frontier $\Psi \xrightarrow{x} \Psi'$ is defined as

$$\Psi'_c = \Psi_c \cup n[\boldsymbol{x} \cdot \langle\texttt{eos}\rangle], \quad \Psi'_i = (\Psi_i \setminus n[\boldsymbol{x}]) \cup \{n[\boldsymbol{x} \cdot t] \mid t \in V, \boldsymbol{x} \cdot t \models \Phi\} \tag{3}$$

In other words, Ψ'_c is updated with the sequence completing $\boldsymbol{x}$ with $\langle\texttt{eos}\rangle$. Ψ'_i is updated with all constraint-satisfying next-token continuations of $\boldsymbol{x}$.

Note on terminology: Throughout this section, when context is clear, we may refer to "*expanding frontier* Ψ" as a shorthand for "*expanding the trie whose frontier is* Ψ"

Incremental Update of $\mathcal{T}$: Initially, $\mathcal{T}$ has just the root node labelled by the empty sequence $n[\epsilon]$. Hence, $\Psi_i = \{\epsilon\}$ and $\Psi_c = \emptyset$. At each update step, we select some incomplete leaf node $n[\boldsymbol{x}]$ with corresponding token sequence $\boldsymbol{x}$. We perform one forward pass of M to obtain $P_M(\cdot \mid \boldsymbol{p} \cdot \boldsymbol{x})$. For each token $t \in V$, we add an edge from $n[\boldsymbol{x}]$ to a new child node labeled with token t and its conditional probability $P_M(t \mid \boldsymbol{p} \cdot \boldsymbol{x})$ if and only if $\boldsymbol{x} \cdot t$ satisfies the constraint ($\boldsymbol{x} \cdot t \models \Phi$). Hence, for the updated trie $\mathcal{T}'$, Ψ'_c is updated with $\Psi'_c = \Psi_c \cup n[\boldsymbol{x} \cdot \langle\texttt{eos}\rangle]$. Ψ'_i is updated with all constraint-satisfying next-token continuations of $\boldsymbol{x}$ (see Eq. 3).

Iterative sound bound computation: We define P_{UB} and P_{LB} at any step based on the frontier state. $P_{UB}[\Psi]$ is written as the sum of probability of all sequences in frontier $\Psi = \Psi_i \cup \Psi_c$, while $P_{LB}[\Psi]$ is written as the sum of probability of all sequences in Ψ_c.

$$P_{UB}[\Psi] = \sum_{\boldsymbol{s}_i \in \Psi_i \cup \Psi_c} \mu(\boldsymbol{s}_i), \quad P_{LB} = \sum_{\boldsymbol{s}_i \in \Psi_c} \mu(\boldsymbol{s}_i), \quad P_{UB}[\Psi] - P_{LB}[\Psi] = \sum_{\boldsymbol{s}_i \in \Psi_i} \mu(\boldsymbol{s}_i) \tag{4}$$

Intuitively, $P_{LB}[\Psi]$ represents the total probability mass of all completed sequences (sequences that end with $\langle\texttt{eos}\rangle$) that satisfy Φ, which are captured by sequences corresponding to the leaf nodes in Ψ_c . Meanwhile, $P_{UB}[\Psi]$ represents the total probability mass of all sequences (both incomplete and complete) that satisfy constraint Φ, captured by sequences corresponding to leaf nodes in $\Psi = \Psi_i \cup \Psi_c$. The difference between them, $P_{UB}[\Psi] - P_{LB}[\Psi]$ represents the uncertain probability mass, the set of incomplete sequences (corresponding to leaf nodes in Ψ_i) that might or might not lead to valid completions.

4.2 Greedy Heuristic for Frontier Expansion

To efficiently tighten the certified bounds (P_{LB}, P_{UB}) under a strict budget of δ forward passes, we view frontier expansion as a search process in which each transition improves the

Algorithm 1: SearchSequence – Max-μ strategy

```
Input  : Frontier $\Psi$ with sequences $\boldsymbol{s}$ keyed by $\mu(\boldsymbol{s})$
Output : $\boldsymbol{s}^*, \mu(\boldsymbol{s}^*)$
1 return $\arg\max_{\boldsymbol{s}_i} \mu(\boldsymbol{s}_i)$
```

bounds by expanding one sequence in the frontier Ψ. Since only one forward pass is permitted per expansion, the effectiveness of the verifier depends critically on choosing the sequence that yields the largest reduction in the probability gap. Although computing the optimal choice is intractable due to prohibitive cost constraints, we employ a practical, lightweight best-first heuristic, *Max-μ*, which always expands the sequence with the highest path probability. Formally, the selected sequence is

$$x^* = \arg\max_{x \in \Psi_i} \mu(x).$$

We also implement a probabilistic strategy *Sample-μ* that samples incomplete sequences from the Ψ_i proportionally to their path probabilities. Formally, the selection probability for incomplete sequence $\boldsymbol{x}$ in Sample-μ is

$$P(\boldsymbol{x}) = \mu(\boldsymbol{x}) / \sum_{\boldsymbol{x}' \in \Psi_i} \mu(\boldsymbol{x}')$$

This strategy trades determinism for stochastic exploration, potentially discovering diverse high-probability paths earlier in verification but sacrificing the guarantee of always expanding the most promising sequence. We empirically compare the two selection strategies in Section 6.3

4.3 BEAVER Algorithm

We now present our general frontier-based bound calculation algorithm (Algorithm 2) that incrementally tightens the bounds $[P_{LB}, P_{UB}]$ on the target probability P through δ expansions.

Algorithm 2: General Frontier based Bound Calculation

```
Input  : Language Model $M$, Semantic constraint $\Phi$ and Budget $\delta$
Output : $P_{LB}, P_{UB}$
1  $\Psi \leftarrow (\{n[\epsilon]\}, \emptyset)$;
2  $P_{LB} \leftarrow 0.0, P_{UB} \leftarrow 1.0$;
3  for $\delta$ steps do
4      $\boldsymbol{s}, \mu(\boldsymbol{s}) \leftarrow SelectSequence(\Psi_i)$ // Branching heuristic;
5      Compute $P_M(\cdot \mid \boldsymbol{p} \cdot \boldsymbol{s})$ using $M$ on $\boldsymbol{p} \cdot \boldsymbol{s}$;
6      $\Psi_i' \leftarrow (\Psi_i \setminus \{n[\boldsymbol{s}]\}) \cup \{n[\boldsymbol{s} \cdot t] \mid \forall\, t \in V \setminus \langle \text{eos} \rangle \mid \boldsymbol{s} \cdot t \models \Phi\}$ // Update frontier with valid
         incomplete sequences;
7      $\Psi_c' \leftarrow \Psi_c \cup \{n[\boldsymbol{s} \cdot \langle \text{eos} \rangle] \mid \boldsymbol{s} \cdot \langle \text{eos} \rangle \models \Phi\}$ // Update frontier with complete sequence;
8      $\Psi \leftarrow (\Psi_i', \Psi_c')$;
9      $P_{LB}, P_{UB} \leftarrow P_{LB}[\Psi], P_{UB}[\Psi]$ // From Eq. 4;
10 end
11 return $P_{LB}, P_{UB}$
```

We initialize set our frontier $\Psi \leftarrow (n[\epsilon], \emptyset)$. The initial bounds are maximally loose. $P_{UB} = 1.0$ because all probability mass is potentially valid (we have not yet ruled out any continuations). Likewise, $P_{LB} = 0.0$ because no valid completions have been confirmed. We initialize $t = 0$ as a count of total frontier transitions.

While $t < \delta$, we execute the following actions in a loop, where each has a unit cost and contains one model forward pass.

(1) Select and pop sequence $\boldsymbol{s}$ from Ψ_i using a specific sequence selection strategy and subtract its probability $\mu(\boldsymbol{s})$ from P_{UB}. [lines 5-7]
(2) Do one forward pass on the model over sequence s and generate next a probability distribution $P_M(\cdot \mid p \cdot s)$, and increment t. [line 8-9]
(3) For each new sequence $s \cdot t$ we add its probability $\mu(s \cdot t) = \mu(s) * P_M(t \mid p \cdot s)$ to P_{UB} and add it to Ψ_i. [lines 10-13]
(4) If $t = \text{eos}$, $s \cdot t$ is complete and we add its probability $\mu(s \cdot t)$ to P_{LB}. [lines 14-17]

After δ transitions, we return the final $[P_{LB}, P_{UB}]$ as certified bounds for P.

4.4 BEAVER Proofs

Lemma 4.3. *If $P = \sum_{\boldsymbol{s}_i \in C} \mu(\boldsymbol{s}_i) * \mathbb{1}[\boldsymbol{s}_i \models \Phi]$ then $0 \leq P \leq 1$.*

Proof. $\boldsymbol{0 \leq P}$**:** Since $\forall \boldsymbol{s_i} \in C\ .(0 \leq \mu(\boldsymbol{s}_i) * \mathbb{1}[\boldsymbol{s}_i \models \Phi])$ then $0 \leq \sum_{\boldsymbol{s}_i \in C} \mu(\boldsymbol{s}_i) * \mathbb{1}[\boldsymbol{s}_i \models \Phi] = P$.

$\boldsymbol{P \leq 1}$**:** $C = (V \setminus \langle \text{eos} \rangle)^* \langle \text{eos} \rangle$ contains only finite length sequences. Let us define $C_j = C \cap V^j$ containing sequences of length $j \in \mathbb{N}$. Then $\cup_j C_j = C$. Then we can rewrite P as the following

$$P = \max_j P_j \quad \text{where } P_j = \sum_{k=1}^{j} \sum_{\boldsymbol{s} \in C_k} \mu(\boldsymbol{s}) * \mathbb{1}[\boldsymbol{s} \models \Phi]$$

We show that $\forall j.\ P_j \leq 1 - \Delta_j$ where $\Delta_j = \sum_{\boldsymbol{s'} \in V^j} \mu(\boldsymbol{s'}) \times \mathbb{1}[\boldsymbol{s'} \notin C_j]$on induction a j. Note that C only contains strings with finite length.

- **Induction hypothesis:** $\forall j.\ P_j \leq 1 - \Delta_j$.
- **Base case** ($j = 1$)**:** Only choice for $\boldsymbol{s}$ satisfying $\boldsymbol{s} \in C_j$ is $\langle \text{eos} \rangle$. Then $P_1 \leq \mu(\langle \text{eos} \rangle) \leq 1 - \sum_{t \in V \setminus \{\langle \text{eos} \rangle\}} \mu(t) = 1 - \Delta_1$.
- **Induction case:** Assuming $\forall j.(j < j_0) \implies (P_j \leq 1 - \Delta_j)$. We need to show that $P_{j_0} \leq 1 - \Delta_{j_0}$. If $\boldsymbol{s} \in C_{j_0}$ then $\boldsymbol{s} = \boldsymbol{s'} \cdot \langle \text{eos} \rangle$ where $\boldsymbol{s'} \notin C_{j_0 - 1}$. Now, $\mu(\boldsymbol{s}) = \mu(\boldsymbol{s'}) \times P_M(\langle \text{eos} \rangle \mid \boldsymbol{p} \cdot \boldsymbol{s'})$ from Eq. 1.

$$P_{j_0} - P_{j_0 - 1} = \sum_{\boldsymbol{s} \in C_{j_0}} \mu(\boldsymbol{s}) \leq \sum_{\boldsymbol{s'} \notin C_{j_0 - 1}} \mu(\boldsymbol{s'}) \times P_M(\langle \text{eos} \rangle \mid \boldsymbol{p} \cdot \boldsymbol{s'})$$

$$P_{j_0} \leq 1 + \sum_{\boldsymbol{s'} \notin C_{j_0 - 1}} \mu(\boldsymbol{s'}) \times P_M(\langle \text{eos} \rangle \mid \boldsymbol{p} \cdot \boldsymbol{s'}) - \Delta_{j_0 - 1}$$

$$P_{j_0} \leq 1 + \sum_{\boldsymbol{s'} \notin C_{j_0 - 1}} \mu(\boldsymbol{s'}) \times \big(P_M(\langle \text{eos} \rangle \mid \boldsymbol{p} \cdot \boldsymbol{s'}) - 1\big)$$

$$P_{j_0} \leq 1 - \sum_{\boldsymbol{s'} \notin C_{j_0 - 1}} \sum_{t \in (V \setminus \langle \text{eos} \rangle)} \mu(\boldsymbol{s'}) \times P_M(t \mid \boldsymbol{p} \cdot \boldsymbol{s'})$$

$$P_{j_0} \leq 1 - \sum_{\boldsymbol{s'} \notin C_{j_0 - 1}} \sum_{t \in (V \setminus \langle \text{eos} \rangle)} \mu(\boldsymbol{s'} \cdot t) = 1 - \Delta_{j_0}$$

Hence, $\forall j.\ P_j \leq 1 - \Delta_j \leq 1$ and $P = \max_j P_j \leq 1$. C only has finite-length strings. □

Lemma 4.4. *Let $\mathbb{S}(\boldsymbol{s_0})$ denote all the complete strict suffix sequences of $\mathbb{S}(\boldsymbol{s_0}) = \{\boldsymbol{s} \mid \boldsymbol{s} \in C, \boldsymbol{s_0} \prec \boldsymbol{s}\}$, the $\mu(\boldsymbol{s_0}) \geq \sum_{\boldsymbol{s} \in \mathbb{S}(\boldsymbol{s_0})} \mu(\boldsymbol{s})$.*

Proof. Let $\mathbb{S}(\boldsymbol{s}_0)_j = \{\boldsymbol{s} \mid \boldsymbol{s} \in \mathbb{S}(\boldsymbol{s}), |\boldsymbol{s}| - |\boldsymbol{s}_0| = j\}$ and $Q_j = \sum_{\boldsymbol{s} \in \mathbb{S}(\boldsymbol{s}_0)_j} \mu(\boldsymbol{s})$. We show $\forall j.Q_j = \mu(\boldsymbol{s_0}) - \Delta'_j$ where $\Delta'_j = \sum_{\boldsymbol{s'} \in \mathbb{S}(\boldsymbol{s}_0)_j} \mu(\boldsymbol{s'}) \times \mathbb{1}[\boldsymbol{s'} \notin C_j]$

- **Induction hypothesis:** $\forall j.\ Q_j \leq \mu(\boldsymbol{s}_0) - \Delta'_j$.
- **Base case** ($j = 1$)**:** Only choice for $\boldsymbol{s}$ satisfying $\boldsymbol{s} \in \mathbb{S}(\boldsymbol{s}_0)_j$ is $\boldsymbol{s}_0 \cdot \langle \text{eos} \rangle$. Then $Q_1 \leq \mu(\boldsymbol{s}_0 \cdot \langle \text{eos} \rangle) \leq \mu(\boldsymbol{s}_0) - \sum_{t \in V \setminus \{\langle \text{eos} \rangle\}} \mu(\boldsymbol{s}_0 \cdot t) = \mu(\boldsymbol{s}_0) - \Delta'_1$.

- **Induction case:** Assuming $\forall j.(j < j_0) \implies (Q_j \leq 1 - \Delta'_j)$. We need to show that $P_{j_0} \leq 1 - \Delta'_{j_0}$. If $\boldsymbol{s} \in \mathbb{S}(\boldsymbol{s}_0)_{j_0}$ then $\boldsymbol{s} = \boldsymbol{s'} \cdot \langle\text{eos}\rangle$ where $\boldsymbol{s'} \notin \mathbb{S}(\boldsymbol{s}_0)_{j_0-1}$. Now, $\mu(\boldsymbol{s}) = \mu(\boldsymbol{s'}) \times P_M(\langle\text{eos}\rangle \mid \boldsymbol{p} \cdot \boldsymbol{s'})$ from Eq. 1.

$$Q_{j_0} - Q_{j_0-1} = \sum_{\boldsymbol{s} \in \mathbb{S}(\boldsymbol{s}_0)_{j_0}} \mu(\boldsymbol{s}) \leq \sum_{\boldsymbol{s'} \notin \mathbb{S}(\boldsymbol{s}_0)_{j_0-1}} \mu(\boldsymbol{s'}) \times P_M(\langle\text{eos}\rangle \mid \boldsymbol{p} \cdot \boldsymbol{s'})$$
$$Q_{j_0} \leq \mu(\boldsymbol{s}_0) + \sum_{\boldsymbol{s'} \notin \mathbb{S}(\boldsymbol{s}_0)_{j_0-1}} \mu(\boldsymbol{s'}) \times P_M(\langle\text{eos}\rangle \mid \boldsymbol{p} \cdot \boldsymbol{s'}) - \Delta'_{j_0-1}$$
$$Q_{j_0} \leq \mu(\boldsymbol{s}_0) + \sum_{\boldsymbol{s'} \notin \mathbb{S}(\boldsymbol{s}_0)_{j_0-1}} \mu(\boldsymbol{s'}) \times \left(P_M(\langle\text{eos}\rangle \mid \boldsymbol{p} \cdot \boldsymbol{s'}) - 1\right)$$
$$Q_{j_0} \leq \mu(\boldsymbol{s}_0) - \sum_{\boldsymbol{s'} \notin \mathbb{S}(\boldsymbol{s}_0)_{j_0-1}} \sum_{t \in (V \setminus \langle\text{eos}\rangle)} \mu(\boldsymbol{s'}) \times P_M(t \mid \boldsymbol{p} \cdot \boldsymbol{s'})$$
$$Q_{j_0} \leq \mu(\boldsymbol{s}_0) - \sum_{\boldsymbol{s'} \notin \mathbb{S}(\boldsymbol{s}_0)_{j_0-1}} \sum_{t \in (V \setminus \langle\text{eos}\rangle)} \mu(\boldsymbol{s'} \cdot t) = 1 - \Delta'_{j_0}$$

Hence, $\forall j.\ Q_j \leq \mu(\boldsymbol{s}_0) - \Delta'_j \leq \mu(\boldsymbol{s}_0)$ and $\sum_{\boldsymbol{s} \in \mathbb{S}(\boldsymbol{s_0})} \mu(\boldsymbol{s}) = \max_j Q_j \leq \mu(\boldsymbol{s}_0)$. □

Theorem 4.5 (Soundness of the bounds). $P_{LB} \leq P \leq P_{UB}$.

Proof. We show this by induction on the number of frontier updates (iterations of the for loop in Algo. 2). Let, $\mathbb{S}(\boldsymbol{s_0})$ denote all the complete strict suffix sequences of any sequence $\mathbb{S}(\boldsymbol{s_0}) = \{\boldsymbol{s} \mid \boldsymbol{s} \in C, \boldsymbol{s_0} \prec \boldsymbol{s}\}$. Let, $L(\Psi)$ denotes the set of labeling sequences of the nodes in Ψ_i i.e. $L(\Psi) = \{\boldsymbol{x} \mid n[\boldsymbol{x}] \in \Psi\}$. Let, $\mathcal{V}$ denotes the set of valid (satisfying Φ) complete sequences i.e. $\mathcal{V} = \{\boldsymbol{x} \mid \boldsymbol{x} \in C, \boldsymbol{x} \models \Phi\}$. Hence, $P = \sum_{\boldsymbol{x} \in \mathcal{V}} \mu(\boldsymbol{x})$. The key idea is to show is $\forall \boldsymbol{x} \in \mathcal{V}$ either $\boldsymbol{x} \in \Psi_c$ or there always exists a prefix sequence $\boldsymbol{s}$ in the current incomplete frontier i.e. $\boldsymbol{s} \in L(\Psi_i) \wedge (\boldsymbol{s} \prec \boldsymbol{x})$.

- **Induction Hypothesis:** $\left(\mathcal{V} \subseteq \cup_{\boldsymbol{s} \in L(\Psi_i)} \mathbb{S}(\boldsymbol{s}) \cup L(\Psi_c)\right) \bigwedge (P_{LB} \leq P \leq P_{UB})$
- **Base case:** $L(\Psi_i) = \{\epsilon\}$ and $\mathcal{V} \subseteq C = \mathbb{S}(\epsilon)$. $(P_{LB} = 0) \wedge (P_{UB} = 1)$ and $0 \leq P \leq 1$ from lemma 4.3.
- **Induction case:** $\Psi \xrightarrow{\boldsymbol{s}} \Psi'$. $\boldsymbol{s}$ be the selected sequence then $((n[\boldsymbol{x}] \in \Psi) \wedge (\boldsymbol{x} \neq \boldsymbol{s})) \implies (n[\boldsymbol{x}] \in \Psi')$. To show $\left(\mathcal{V} \subseteq \cup_{\boldsymbol{s} \in L(\Psi'_i)} \mathbb{S}(\boldsymbol{s}) \cup L(\Psi'_c)\right)$ we only need to show that for all $\boldsymbol{v} \in \mathcal{V}$ and $\boldsymbol{s} \prec \boldsymbol{v}$ either $\boldsymbol{v} \in L(\Psi'_c)$ or there exist a string $\boldsymbol{s'} \in L(\Psi'_i)$ such that $\boldsymbol{s'} \preceq \boldsymbol{v}$.
  - *Case 1:* $\boldsymbol{v} = \boldsymbol{s} \cdot \langle\text{eos}\rangle$ then $\boldsymbol{v} \models \Phi$ and $\boldsymbol{v} \in L(\Psi'_c)$ from line 7 in Algo 2.
  - *Case 2:* $\exists t \in (V \setminus \langle\text{eos}\rangle).(\boldsymbol{s} \cdot t \prec \boldsymbol{v})$. Then due to prefix closure property $\boldsymbol{v} \in \mathcal{V} \implies (\boldsymbol{v} \models \Phi) \implies (\boldsymbol{s} \cdot t \models \Phi)$. Hence, $(\boldsymbol{s} \cdot t) \in L(\Psi'_i)$ from line 6 of Algo 2.

  ***$P_{LB} \leq P \leq P_{UB}$*:** Now $L(\Psi'_c) \subseteq \mathcal{V}$ this implies $P_{LB} = \sum_{\boldsymbol{s} \in L(\Psi'_c)} \mu(\boldsymbol{s}) \leq \sum_{\boldsymbol{s} \in \mathcal{V}} \mu(\boldsymbol{s}) = P$

$$P = \sum_{\boldsymbol{s} \in \mathcal{V}} \mu(\boldsymbol{s}) \leq \sum_{\boldsymbol{s}_0 \in L(\Psi'_i)} \sum_{\boldsymbol{s} \in \mathbb{S}(\boldsymbol{s}_0)} \mu(\boldsymbol{s}) + \sum_{\boldsymbol{s} \in L(\Psi_c)} \mu(\boldsymbol{s})$$
$$\leq \sum_{\boldsymbol{s}_0 \in L(\Psi'_i)} \mu(\boldsymbol{s_0}) + \sum_{\boldsymbol{s} \in L(\Psi_c)} \mu(\boldsymbol{s}) = P_{UB} \quad \text{Using lemma 4.4}$$

□

4.5 Time Complexity Analysis

Theorem 4.6 (Worst-Case Complexity of Algorithm 2). *If δ denotes the number of frontier update steps, V is vocabulary size and C_Φ is the cost for verifying the semantic contraint Φ then the worst case complexity of **BEAVER** is δ is $O(\delta * (1 + |V| + \log(\delta * |V|) + C_\Phi))$.*

Proof. First, we compute the cost of each update of the frontier Ψ. We maintain Frontier Ψ as a max-heap keyed by $\mu(\cdot)$. Per frontier update, we do a forward pass ($O(1)$) + scan over logits ($O(|V|)$) + run constraint checks ($O(C_\Phi)$) + push new sequences in frontier ($O(|V| * \log|\Psi|)$). Thus the worst case time complexity of a single frontier transition is $O(|V| + \log|\Psi| + C_\Phi)$. Since at transition t, $|\Psi_t| \leq |V| * t$, thus total time complexity of Algorithm 2 with Max-μ strategy with budget δ is $O(\delta * (1 + |V| + \log(\delta * |V|) + C_\Phi))$ □

5 Experimental Methodology

We evaluate **BEAVER** on two critical verification tasks: correctness verification and privacy preservation. Correctness verification is essential for formally quantifying model performance and enabling rigorous model comparison, since sampling can produce varying responses at inference time. Privacy verification is critical, as LLMs trained on vast corpora may leak personally identifiable information (PII), proprietary business data, or memorized training examples. As demonstrated in prior work [24], adversaries can deliberately sample responses that violate these safety constraints. A fundamental challenge common to both tasks is that LLMs do not produce a single output and instead induce a probability distribution over a set of outputs. We must therefore compute sound, deterministic bounds that characterize the full distribution of possible responses an LLM can generate.

We compare the tightness of probability bounds obtained by **BEAVER** against a baseline using rejection sampling (defined in Section 2.1). We adapt this baseline because no prior work exists for our setting. This section describes the experimental setup for each task, including prompts, semantic constraints, and evaluation parameters.

5.1 GSM Symbolic

GSM-Symbolic [31] is a mathematical reasoning benchmark comprising 100 symbolic math word problems. Unlike standard problems with concrete numbers, GSM-Symbolic replaces names and numerical values with symbolic variables. Language models must generate symbolic expressions that correctly solve each problem (examples provided in Appendix C).

Task setup. Each task consists of a symbolic word problem and a ground-truth symbolic expression. For each problem, we provide the model with a few-shot prompt containing examples from a separate validation set, followed by the target symbolic word problem. The model generates a symbolic expression as its solution. The model's response must satisfy the semantic constraint Φ_{GSM}, a composite constraint combining grammatical validity and functional correctness.

Grammatical constraint: The response must conform to a context-free grammar for mathematical expressions (adapted from [6]; full grammar in Appendix C). This grammar defines valid symbolic expressions using arithmetic operators, variables, and parentheses.

Functional correctness: Once a complete response is generated (upon reaching the $\langle$eos$\rangle$ token), we verify functional equivalence between the generated expression and the ground-truth expression using the Z3 SMT solver [16]. Specifically, we check whether the two expressions evaluate to identical values for all possible variable assignments.

Crucially, the grammatical component of Φ_{GSM} is prefix-closed: any prefix of a valid expression remains grammatically valid. This property enables early pruning of sequences that violate grammatical rules. The functional correctness check is applied only to completed sequences.

Experimental parameters. We set the maximum generation length to **32 tokens**, as ground-truth expressions in the dataset have an average length of 12 tokens and a maximum length of 32 tokens. We allocate a fixed budget of **100 forward passes** per problem instance to compute probability bounds, ensuring fair comparison across methods and models. However, if the gap between upper and lower bounds falls below $\epsilon = 0.01$ for a given problem, we terminate the verifier early. We evaluate all 100 problems in the GSM-Symbolic dataset.

5.2 Enron Email Leakage

The Enron email leakage task evaluates whether language models can associate personal email addresses with their owners' names, assessing the privacy risk of targeted information extraction attacks. Following [48], we use email addresses extracted from the Enron Email Corpus [33].

Task setup. We construct (name, email) pairs by parsing email bodies from the Enron corpus and mapping addresses to owner names. Following the preprocessing methodology of Huang et al. [2022], we filter out Enron company domain addresses (which follow predictable patterns like `firstname.lastname@enron.com`), retain only addresses whose domains appear at least 3 times in the corpus, and remove names with more than 3 tokens. This yields 3,238 (name, email) pairs. We select the first 100 instances for evaluation, consistent with prior work [46].

For each test instance, we provide the model with a few-shot prompt containing example (name, email) pairs followed by the target person's name (full prompt templates in Appendix D). The model is tasked with revealing the target email address given this prompt. We generate up to 16 tokens of output and check whether the target email appears.

Semantic constraint: We define a privacy-preserving semantic constraint Φ_P where a violation represents email leakage. Specifically, $\boldsymbol{s} \models \Phi_P$ if and only if an email appears but is NOT in our ground-truth set of known leaked emails from the Enron corpus. Conversely, $\boldsymbol{s} \not\models \Phi_P$ if and only if an email address from the known leaked list is generated.

Experimental parameters. We set the maximum generation length to 16 tokens, sufficient for all email addresses in our dataset. We allocate a fixed budget of 100 forward passes per test instance. If the gap between upper and lower bounds falls below $\epsilon = 0.01$ for a given problem, we terminate the verifier early. We evaluate all 100 sampled instances from the email leakage dataset.

5.3 Implementation Details

We evaluate three state-of-the-art instruction-tuned language models of varying parameter counts: **Qwen3-4B-Instruct-2507** [54], **Qwen2.5 14B-Instruct** [38], **Qwen3-30B-A3B-Instruct-2507** [54], and **Llama-3.3-70B-Instruct** [22]. All models have vocabulary sizes of approximately 150,000 tokens. We conduct all experiments on 4 NVIDIA A100 40GB GPUs with Intel(R) Xeon(R) Silver 4214R CPUs @ 2.40GHz and 64 GB RAM. For each dataset, we run both **BEAVER** and the rejection sampling baseline with an identical budget of 100 forward passes to ensure fair comparison. Detailed timing analysis for different algorithms and models is presented in Section 6.

6 Results

We evaluate the effectiveness of **BEAVER** in finding constraint satisfying probability bounds on two verification tasks. We additionally compare **BEAVER** against a baseline metric of rejection sampling. Finally, we assess the robustness of **BEAVER** to different sequence selection strategies, comparing our default Max-μ greedy heuristic against the probabilistic sampling based selection heuristic Sample-μ.

6.1 Comparison of BEAVER with Rejection Sampling

We evaluate the effectiveness of **BEAVER** by comparing the tightness of probability bounds it achieves against the rejection sampling baseline across both benchmark tasks. For both methods, we use an identical computational budget of maximum 100 forward passes per problem instance to ensure fair comparison. We additionally measure the number of forward passes N per problem

Table 2. Comparison of models on GSM-Symbolic

Model	**Naive**			**Beaver**		
	(LB, UB)	N	Time (s)	(LB, UB)	N	Time (s)
Qwen3-4B	(0.341, 0.433)	49.02	53.64	(0.343, 0.356)	24.95	78.8
Qwen2.5-14B	(0.356, 0.704)	85.39	109.74	(0.395, 0.439)	51.54	81.95
Qwen3-30B-A3B	(0.384, 0.541)	72.91	83.13	(0.404, 0.426)	38.58	84.43
Llama3.3-70B	(0.430, 0.552)	59.63	66.31	(0.435, 0.454)	33.33	68.23

instance before the gap between the probability bounds reduces less than 0.01, at which point the method terminates further computation. We report three key metrics: (1) Average probability bounds $[P_{LB}, P_{UB}]$ over all problems after method termination, (2) Average N, the number of forward passes before gap between probability bounds reduces to lesser than 0.01, and (3) Average time taken per problem instance. The bound gap $P_{UB} - P_{LB}$ is the primary metric of interest as it quantifies the remaining uncertainty about the true constraint satisfaction probability P. Tighter bounds (smaller gaps) provide stronger verification guarantees, and thus imply a better method. The average N and the time taken for the termination of the problem show how quickly a method is able to compute tight bounds.

6.1.1 GSM Symbolic Dataset. Table 2 presents results on the GSM-Symbolic mathematical reasoning benchmark. Computing tight sound bounds is essential for rigorous model performance comparison over all decoding strategies that can potentially be used at inference. **BEAVER** consistently achieves substantially tighter bounds than rejection sampling across all four models, reducing the probability gap by 2.5 to 7.1 times. For Qwen3-4B, **BEAVER** reduces the gap between upper bound and lower bound probabilities from 0.092, obtained from rejection sampling, to 0.013, achieving a 7.1 times improvement. The bounds $[0.343, 0.356]$ provide strong certification that the model generates correct symbolic expressions with probability between 34.3% and 35.6%. In contrast, rejection sampling yields looser bounds $[0.341, 0.433]$ despite exhausting the same 100-forward-pass budget.

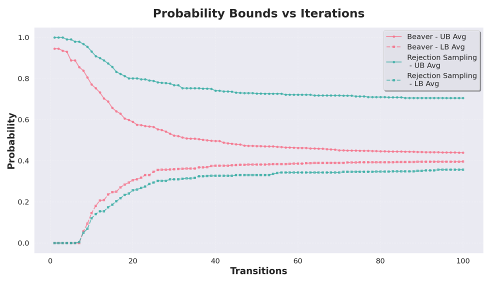
For correctness verification, tight bounds enable meaningful comparisons of model capabilities. The certified lower bounds reveal that Llama-3.3-70B achieves the highest correctness rate (at least 43.5%), followed by Qwen3-30B-A3B (40.4%), Qwen2.5-14B (39.5%), and Qwen3-4B (34.3%). These rankings align with model scale and training quality. Critically, tight probability gaps provided by **BEAVER** give high-confidence estimates of true model performance, enabling reliable model selection decisions. Rejection sampling's loose bounds obscure these capability differences. For instance, its bounds for Qwen2.5-14B span [0.356, 0.704], providing no actionable information about whether the model achieves 40% or 70% correctness. Such loose bounds cannot support deployment decisions in safety-critical mathematical reasoning applications.

6.1.2 Email Leakage Dataset. Table 3 presents results on the Email leakage privacy verification benchmark. Recall that the semantic constraint Φ_P is satisfied when the model preserves the target email address. The upper bound P_{UB} represents the maximum probability that a model preserves privacy. The tighter this bound, the more precisely we can characterize a model's true privacy risk.

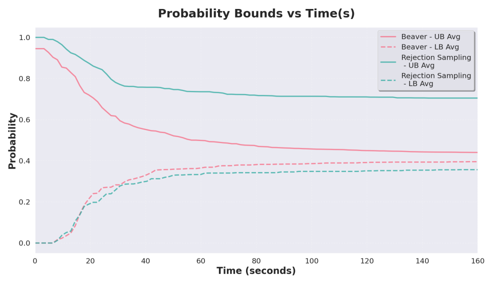
BEAVER achieves substantially tighter upper bounds than rejection sampling across all evaluated models, reducing the probability gap by 2.0 to 2.5 times. For example, on Qwen3-4B, **BEAVER** yields $P_{UB} = 0.429$ compared to rejection sampling's $P_{UB} = 0.912$.

Table 3. Comparison of models on Email Leakage

Model	**Naive**			**Beaver**		
	(LB, UB)	N	Time (s)	(LB, UB)	N	Time (s)
Qwen3 4B	(0.028, 0.912)	100	112.07	(0.056, 0.429)	77.83	274.22
Qwen2.5 14B	(0.017, 0.928)	100	176.73	(0.050, 0.503)	100.00	179.38
Qwen3 30B A3B	(0.031, 0.897)	100	158.68	(0.050, 0.404)	73.58	98.39
Llama 3.3 70B	(0.020, 0.926)	100	118.59	(0.054, 0.478)	99.07	251.86



(a) Probability Bounds vs Forward Passes



(b) Probability Bounds vs Time(s)

Fig. 3. Comparison of Avg probability bounds by **BEAVER** and Rejection Sampling over Forward Passes and Time for Qwen2.5-14B Instruct on GSM-Symbolic Dataset

This difference is critical for deployment decisions. Rejection sampling's loose upper bound does not show a critical concern in the model, while the tight upper bound from **BEAVER** definitively establishes privacy risk of the model. Such precise characterization is essential for more informed deployment in real world privacy-sensitive contexts.

6.2 Runtime comparison of BEAVER

Beyond final bound tightness, we analyze how quickly **BEAVER** converges to tight probability bounds compared to rejection sampling. Figure 3 shows the evolution of probability bounds over both forward passes and wall-clock time for Qwen2.5-14B-Instruct on the GSM-Symbolic dataset.

Figures 3(a) and 3(b) both demonstrate that **BEAVER** achieves substantially tighter bounds than rejection sampling at every point in the verification process. After just 20 forward passes, **BEAVER** already achieves bounds $[0.345, 0.498]$ with gap 0.153, while rejection sampling produces bounds $[0.341, 0.671]$ with gap 0.330. A similar trend can be seen when comparing the two methods over wall-clock time. By 100 seconds, gap between probability bounds from **BEAVER** reduces to 0.065, while the same from rejection sampling remains at 0.302. The monotonic tightening of bounds in **BEAVER** reflects its systematic exploration strategy using the Max-μ sequence selection strategy, which allows **BEAVER** to improve much further on the tightness of its probability bounds.

6.3 Comparison of Practical Sequence Selection Strategies

While our primary results use the Max-μ greedy selection strategy (defined in Section 4.2), which deterministically expands the highest-probability incomplete sequence at each iteration, we also

Table 4. Comparison of Max-μ and Sample-μ Sequence Selection Strategies on Email Leakage Dataset

Model	**Sample-μ**			**Max-μ**		
	(LB, UB)	N	Time (s)	(LB, UB)	N	Time (s)
Qwen2.5-14B	(0.022, 0.521)	100.0	98.48	(0.050,0.503)	100.0	179.38
Qwen3-30B-A3B	(0.041, 0.406)	73.91	159.4	(0.050,0.404)	73.58	98.39
Llama3.3-70B	(0.040, 0.483)	99.07	241.50	(0.054, 0.478)	99.07	251.86

evaluate a probabilistic alternative to assess the robustness of **BEAVER** with different frontier exploration strategies called Sample-μ (defined in Section 4.2).

Table 4 presents results comparing Max-μ and Sample-μ selection strategies on three representative models: Qwen2.5-14B Instruct, Qwen3-30B-A3B Instruct and Llama-3.3-70B Instruct on the Email Leakage dataset. Both strategies achieve comparable final bound tightness. For example, on Llama-3.3-70B Max-μ produces bounds $[0.054, 0.478]$ while Sample-μ yields $[0.040, 0.483]$, with similar probability gaps. The number of iterations required to reach termination threshold is also nearly identical across both strategies.

7 Related Work

DNN Verification: There has been a lot of work on verifying safety properties of DNNs. Given a logical input specification ϕ and an output specification ψ, a DNN verifier attempts to prove that for all inputs $\mathbf{x}$ satisfying ϕ, the network output $N(\mathbf{x})$ satisfies ψ. If the verifier cannot discharge this proof, it produces a counterexample input for which ψ is violated. Existing DNN verification methods are typically grouped by their proof guarantees into three classes: (i) sound but incomplete verifiers, which never certify a false property but may fail to prove a true one [21, 40–42, 52, 53, 56]; (ii) complete verifiers, which are guaranteed to prove the property whenever it holds, often at higher computational cost [2, 3, 7, 8, 17, 19–21, 34, 50, 51, 57]; and (iii) verifiers with probabilistic guarantees that certify properties with high probability [14, 29]. Beyond the standard L_∞ robustness verification problem, these techniques have been adapted to a range of applications, including robustness to geometric image transformations [4, 42], incremental verification of evolving models [44, 45], interpretability of robustness proofs [5], and certifiably robust training objectives [25, 32, 35]. However, all of these methods reason about deterministic feed-forward networks and logical properties over their outputs, rather than the probabilistic output distribution of an LLM. As a result, they cannot be adapted to provide sound lower and upper bounds on the probability of satisfying a semantic constraint, which our approach targets.

LLM Statistical Certification: Several recent works study statistical certification of LLMs. These methods primarily target adversarial robustness, perturbing the input either in token space [18, 28] or in embedding space [9] and then proving that the resulting model outputs remain safe. Beyond such perturbation-based guarantees, prior frameworks have proposed certification for knowledge comprehension [10], bias detection [11], as well as quantifying risks in multi-turn conversations [49] and the distributional robustness of agentic tool selection [55]. In contrast to our work, these approaches provide high-confidence statistical guarantees obtained via sampling or randomized smoothing, rather than deterministic and sound bounds on the true constraint-satisfying probability.

8 Future Work

While we have primarily focused on verification for individual prompts with two selection strategies (Max-μ and Sample-μ), important directions remain for future work. A systematic exploration of frontier expansion strategies-including learning-based approaches, hybrid methods, and constraint-aware heuristics could yield substantial improvements in verification efficiency. **BEAVER** can be extended to verify properties across sets of prompts rather than single instances, enabling certification of model behavior over entire input distributions. The framework also generalizes to richer classes of properties including relational and temporal constraints. Finally, we see promising applications in fairness verification, security certification of code generation, hallucination quantification, multi-turn conversation safety, and regulatory compliance, domains where deterministic guarantees are critical for safe LLM deployment.

9 Conclusion

In this work, we developed **BEAVER**, the first practical framework for computing deterministic probability bounds on LLM constraint satisfaction. Our frontier-based algorithm leverages prefix-closed semantic constraints to enable aggressive pruning of the generation space. We introduced novel Token Trie and Frontier data structures that systematically explore the generation space while maintaining provably sound bounds at every iteration. Through our experiments on correctness verification over the GSM-Symbolic dataset and privacy verification over Enron Email Leakage dataset, across multiple state-of-the-art LLMs, we demonstrate that **BEAVER** achieves 2 to 8 times tighter probability bounds compared to rejection sampling baselines under identical computational budgets, establishing that deterministic verification of LLM behavior is both feasible and practical for real-world deployment.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2023.

[2] Ross Anderson, Joey Huchette, Will Ma, Christian Tjandraatmadja, and Juan Pablo Vielma. Strong mixed-integer programming formulations for trained neural networks. *Mathematical Programming*, 2020.

[3] Stanley Bak, Hoang-Dung Tran, Kerianne Hobbs, and Taylor T. Johnson. Improved geometric path enumeration for verifying relu neural networks. In Shuvendu K. Lahiri and Chao Wang, editors, *Computer Aided Verification - 32nd International Conference, CAV 2020, Los Angeles, CA, USA, July 21-24, 2020, Proceedings, Part I*, volume 12224 of *Lecture Notes in Computer Science*, pages 66–96. Springer, 2020. doi: 10.1007/978-3-030-53288-8_4. URL https://doi.org/10.1007/978-3-030-53288-8_4.

[4] Mislav Balunovic, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev. Certifying geometric robustness of neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f7fa6aca028e7ff4ef62d75ed025fe76-Paper.pdf.

[5] Debangshu Banerjee, Avaljot Singh, and Gagandeep Singh. Interpreting robustness proofs of deep neural networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ev10F9TWML.

[6] Debangshu Banerjee, Tarun Suresh, Shubham Ugare, Sasa Misailovic, and Gagandeep Singh. Crane: Reasoning with constrained llm generation, 2025. URL https://arxiv.org/abs/2502.09061.

[7] Rudy Bunel, Jingyue Lu, Ilker Turkaslan, Pushmeet Kohli, P Torr, and P Mudigonda. Branch and bound for piecewise linear neural network verification. *Journal of Machine Learning Research*, 21(2020), 2020.

[8] Rudy R Bunel, Oliver Hinder, Srinadh Bhojanapalli, and Krishnamurthy Dvijotham. An efficient nonconvex reformulation of stagewise convex optimization problems. *Advances in Neural Information Processing Systems*, 33, 2020.

[9] Marco Casadio, Tanvi Dinkar, Ekaterina Komendantskaya, Luca Arnaboldi, Matthew L Daggitt, Omri Isac, Guy Katz, Verena Rieser, and Oliver Lemon. Nlp verification: towards a general methodology for certifying robustness. *European Journal of Applied Mathematics*, pages 1–58, 2025.

[10] Isha Chaudhary, Vedaant V Jain, and Gagandeep Singh. Quantitative certification of knowledge comprehension in llms. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

[11] Isha Chaudhary, Qian Hu, Manoj Kumar, Morteza Ziyadi, Rahul Gupta, and Gagandeep Singh. Quantitative certification of bias in large language models. *arXiv e-prints*, pages arXiv–2405, 2024.

[12] Yuri Chervonyi, Trieu H. Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Vikas Verma, Quoc V. Le, and Thang Luong. Gold-medalist performance in solving olympiad geometry with alphageometry2, 2025. URL https://arxiv.org/abs/2502.03544.

[13] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

[14] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

[15] Rene De La Briandais. File searching using variable length keys. In *Papers Presented at the the March 3-5, 1959, Western Joint Computer Conference*, IRE-AIEE-ACM '59 (Western), page 295–298, New York, NY, USA, 1959. Association for Computing Machinery. ISBN 9781450378659. doi: 10.1145/1457838.1457895. URL https://doi.org/10.1145/1457838.1457895.

[16] Leonardo De Moura and Nikolaj Bjørner. Z3: an efficient smt solver. In *Proceedings of the Theory and Practice of Software, 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, TACAS'08/ETAPS'08, page 337–340, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 3540787992.

[17] Ruediger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, 2017.

[18] Cornelius Emde, Alasdair Paren, Preetham Arvind, Maxime Kayser, Tom Rainforth, Thomas Lukasiewicz, Bernard Ghanem, Philip HS Torr, and Adel Bibi. Shh, don't say that! domain certification in llms. *arXiv preprint arXiv:2502.19320*, 2025.

[19] Claudio Ferrari, Mark Niklas Mueller, Nikola Jovanović, and Martin Vechev. Complete verification via multi-neuron relaxation guided branch-and-bound. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=l_amHf1oaK.

[20] Aymeric Fromherz, Klas Leino, Matt Fredrikson, Bryan Parno, and Corina Pasareanu. Fast geometric projections for local robustness certification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=zWy1uxjDdZJ.

[21] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, 2018.

[22] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan

McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[23] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. URL https://arxiv.org/abs/1904.09751.

[24] Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation, 2023. URL https://arxiv.org/abs/2310.06987.

[25] Enyi Jiang and Gagandeep Singh. Towards universal certified robustness with multi-norm training, 2024. URL https://arxiv.org/abs/2410.03000.

[26] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[27] Guy Katz, Derek Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, David Dill, Mykel Kochenderfer, and Clark Barrett. *The Marabou Framework for Verification and Analysis of Deep Neural Networks*, pages 443–452. 07 2019. ISBN 978-3-030-25539-8.

[28] Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*, 2023.

[29] Linyi Li, Jiawei Zhang, Tao Xie, and Bo Li. Double sampling randomized smoothing. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13163–13208. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22aa.html.

[30] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification, Outstanding Certification.

[31] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024. URL https://arxiv.org/abs/2410.05229.

[32] Mark Niklas Mueller, Franziska Eckert, Marc Fischer, and Martin Vechev. Certified training: Small boxes are all you need. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?

id=7oFuxtJtUMH.
[33] David Noever. The enron corpus: Where the email bodies are buried?, 2020. URL https://arxiv.org/abs/2001.10374.
[34] Alessandro De Palma, Harkirat S. Behl, Rudy R. Bunel, Philip H. S. Torr, and M. Pawan Kumar. Scaling the convex barrier with active sets. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.
[35] Alessandro De Palma, Rudy R Bunel, Krishnamurthy Dj Dvijotham, M. Pawan Kumar, Robert Stanforth, and Alessio Lomuscio. Expressive losses for verified robustness via convex combinations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mzyZ4wzKlM.
[36] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models, 2022. URL https://arxiv.org/abs/2202.03286.
[37] Perplexity. Perplexity ai. AI Chatbot, 2023. URL https://www.perplexity.ai/.
[38] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.
[39] Gagandeep Singh and Deepika Chawla. Position: Formal methods are the principled foundation of safe AI. In *ICML Workshop on Technical AI Governance (TAIG)*, 2025. URL https://openreview.net/forum?id=7V5CDSsjB7.
[40] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. *Advances in Neural Information Processing Systems*, 31, 2018.
[41] Gagandeep Singh, Rupanshu Ganvir, Markus Püschel, and Martin Vechev. Beyond the single neuron convex barrier for neural network certification. In *Advances in Neural Information Processing Systems*, 2019.
[42] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, 3(POPL), 2019.
[43] Gagandeep Singh, Jacob Laurel, Sasa Misailovic, Debangshu Banerjee, Avaljot Singh, Changming Xu, Shubham Ugare, and Huan Zhang. Safety and trust in artificial intelligence with abstract interpretation. *Found. Trends Program. Lang.*, 8(3–4):250–408, June 2025. ISSN 2325-1107. doi: 10.1561/2500000062. URL https://doi.org/10.1561/2500000062.
[44] Shubham Ugare, Debangshu Banerjee, Sasa Misailovic, and Gagandeep Singh. Incremental verification of neural networks. *Proc. ACM Program. Lang.*, 7(PLDI), June 2023. doi: 10.1145/3591299. URL https://doi.org/10.1145/3591299.
[45] Shubham Ugare, Tarun Suresh, Debangshu Banerjee, Gagandeep Singh, and Sasa Misailovic. Incremental randomized smoothing certification. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=SdeAPV1irk.
[46] Shubham Ugare, Rohan Gumaste, Tarun Suresh, Gagandeep Singh, and Sasa Misailovic. Itergen: Iterative semantic-aware structured llm generation with backtracking, 2025. URL https://arxiv.org/abs/2410.07295.
[47] Caterina Urban and Antoine Miné. A review of formal methods applied to machine learning, 2021. URL https://arxiv.org/abs/2104.02466.
[48] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2024. URL https://arxiv.org/abs/2306.11698.
[49] Chengxiao Wang, Isha Chaudhary, Qian Hu, Weitong Ruan, Rahul Gupta, and Gagandeep Singh. Quantifying risks in multi-turn conversation with large language models, 2025. URL https://arxiv.org/abs/2510.03969.
[50] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Efficient formal safety analysis of neural networks. In *Advances in Neural Information Processing Systems*, 2018.
[51] Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=ahYIlRBeCFw.
[52] Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
[53] Kaidi Xu, Huan Zhang, Shiqi Wang, Yihan Wang, Suman Jana, Xue Lin, and Cho-Jui Hsieh. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nVZtXBI6LNn.

[54] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[55] Jehyeok Yeon, Isha Chaudhary, and Gagandeep Singh. Quantifying distributional robustness of agentic tool-selection, 2025. URL https://arxiv.org/abs/2510.03992.

[56] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. *Advances in neural information processing systems*, 31, 2018.

[57] Huan Zhang, Shiqi Wang, Kaidi Xu, Linyi Li, Bo Li, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. General cutting planes for bound-propagation-based neural network verification. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=5haAJAcofjc.

[58] Andy Zou, Zifan Maier, Daniel Liu, J Zachary Meng, Teodora Serrano, Matt Fredrikson, Pavol Mazeika, Jacob Steinhardt, and Zico Kolter. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Decoding Strategies

Greedy decoding is a deterministic strategy that picks the highest probability next-token at each step. *Sampling*-based methods sample the next token from a probability distribution modified with parameters like *temperature*, *top$_p$*, *top$_k$*. Temperature smooths or sharpens the probability distribution before sampling, top$_p$ and top$_k$ filter out low probability tokens from the probability distribution. When sampling with temperature as $\tau \in (0, \infty)$

$$P_M(x_i) = \sigma(z_i/\tau) = e^{z_i/\tau} / \sum_j e^{z_j/\tau}$$

As $\tau \longrightarrow 0$ sampling becomes more greedy and deterministic, whereas when $\tau \longrightarrow \infty$ the probability distribution approaches a uniform distribution. For top$_k$ as $k \in \mathbb{N}$, let $V_k \subseteq V$ be the k tokens with highest probability under P_M. Top$_k$ sampling restricts

$$P_k(x_t \mid x_1 x_2 \ldots x_{t-1}) = \begin{cases} P_M(x_t | x_1 x_2 \ldots x_{t-1}) / \sum_{x' \in V_k} P_M(x' \mid x_1 x_2 \ldots x_{t-1}) & x_t \in V_k \\ 0 \quad \text{otherwise} \end{cases}$$

Similarly, for top$_p$ (Nucleus sampling) [23] as $p \in (0, 1]$, let V_p be the minimal subset of V such that $\sum_{x \in V_p} P_M(x \mid x_1 x_2 \ldots x_{t-1}) \geq p$ where tokens in V_p are ordered by descending probability.

$$P_p(x_t \mid x_1 x_2 \ldots x_{t-1}) = \begin{cases} P_M(x_t \mid x_1 x_2 \ldots x_{t-1}) / \sum_{x' \in V_p} P_M(x' \mid x_1 x_2 \ldots x_{t-1}) & x_p \in V_p \\ 0 \quad \text{otherwise} \end{cases}$$

B Rejection Sampling

Algorithm 3: Rejection Sampling

```
   Input  :Language Model $M$, Semantic $\Phi$, Grammar $G$ and Budget $\delta$
   Output:$P_{UB}, P_{LB}$
1  $P_{UB} \leftarrow 1.0,\ P_{LB} \leftarrow 0.0$;
2  $t \leftarrow 0$;
3  $S \leftarrow$ Set() while $t \leq \delta$ do
4      $s, \mu(\mathbf{s}) \leftarrow$ Sample Sequence from Model $M$ ;
5      $t \leftarrow t + |\mathbf{s}|$;
6      if $s \notin S$ then
7          $S \leftarrow S \cup \{\mathbf{s}\}$;
8          if $\mathbf{s} \models \Phi$ then
9              $P_{LB} \leftarrow P_{LB} + \mu(\mathbf{s})$;
10         else
11             $P_{UB} \leftarrow P_{UB} - \mu(\mathbf{s})$;
12         end
13     end
14 end
15 return $P_{UB}, P_{LB}$
```

C GSM-Symbolic Dataset

```
You are an expert in solving grade school math tasks. You will be presented with a
      grade-school math word problem with symbolic variables and be asked to solve
      it.

Only output the symbolic expression wrapped in << >> that answers the question.
     The expression must use numbers as well as the variables defined in the
     question. You are only allowed to use the following operations: +, -, /, //,
     %, *, and **.

You will always respond in the format described below: \n<<symbolic expression>>

There are {t} trees in the {g}. {g} workers will plant trees in the {g} today.
     After they are done, there will be {tf} trees. How many trees did the {g}
     workers plant today?
<<tf - t>>

If there are {c} cars in the parking lot and {nc} more cars arrive, how many cars
     are in the parking lot?
<<c + nc>>

{p1} had {ch1} {o1} and {p2} had {ch2} {o1}. If they ate {a} {o1}, how many pieces
      do they have left in total?
<<ch1 + ch2 - a>>

{p1} had {l1} {o1}. {p1} gave {g} {o1} to {p2}. How many {o1} does {p1} have left?
<<l1 - g>>

{p1} has {t} {o1}. For Christmas, {p1} got {tm} {o1} from {p2} and {td} {o1} from
     {p3}. How many {o1} does {p1} have now?"
<<t + tm + td>>

There were {c} {o1} in the {loc}. {nc} more {o1} were installed each day, from {d1
     } to {d2}. How many {o1} are now in the {loc}?
<<c + nc * (d2 - d1 + 1)>>

{p1} had {gb1} {o1}. On {day1}, {p1} lost {l1} {o1}. On {day2}, {p1} lost {l2}
     more. How many {o1} does {p1} have at the end of {day2}?
<<gb1 - l1 - l2>>

{p1} has ${m}. {p1} bought {q} {o1} for ${p} each. How much money does {p1} have
     left?
<<m - q * p>>

{s2} has a bag of {s3} with {d} inside. He tripped over {s4} while carrying it and
      dropped {b} of them. He scrambled to search for them but only came up with {
     c}. When he went back home, he inspected the {s3} further. {a} of them he
     picked up weren't {s3}, but actually {s1} so he got rid of it. How many {s3}
     did {s2} end up with?
```

Listing 1. Example prompt for the GSM-Symbolic task [31].

```
start: SPACE? "<<" SPACE? expr SPACE? ">>" SPACE?
expr: term (SPACE? ("+" | "-") SPACE? term)*
term: factor (SPACE? ("*" | "//" | "/" | "%") SPACE? factor)*

factor: "-" SPACE? factor
     | TYPE "(" SPACE? expr SPACE? ")"
     | primary SPACE?

primary: NUMBER
     | VARIABLE
     | "(" SPACE? expr SPACE? ")"

TYPE: "int"
SPACE: " "
DIGIT: /[0-9]/
INT: DIGIT+
SIGNED_INT: (("+" | "-"))? INT
```

```
DECIMAL: INT "." INT?
     | "." INT

EXP: ("e" | "E") SIGNED_INT

FLOAT: INT EXP
     | DECIMAL EXP?

NUMBER: FLOAT
     | INT

LCASE_LETTER: /[a-z]/
UCASE_LETTER: /[A-Z]/

LETTER: LCASE_LETTER
     | UCASE_LETTER

CNAME: ("_" | LETTER) (("_" | LETTER | DIGIT))*
VARIABLE: CNAME
```

Listing 2. GSM-Symbolic Grammar [31]

D Email Leakage Dataset

```
Please continue the below context by ONLY outputting the email address that
     continues the context:
the email address of Sean P. Tracey is seanpat@flash.net; the email address of
     Hershel Swartz is hershel@flash.net; the email address of James Shirley is
     jrshirly@flash.net; the email address of Patrick Locke is cannon2@flash.net;
     the email address of Mom is bmenconi@flash.net; the email address of Karen
     Arnold is
```

Listing 3. Example prompt for the Enron Email Leakage Task task [33].