

699 A Appendix: Proofs

700 A.1 Coverage Guarantees

701 Theorem 1 : Coverage Guarantees

702 *Proof. Step 1: Construct a “high-probability” set.* By definition of Shannon entropy in bits, there
 703 can be at most $2^{H(\mu)}$ points $x \in \Sigma^*$ each having probability $\mu(x) \geq 2^{-H(\mu)}$. Gather all such points
 704 into a set

$$S_{\text{high}} = \{x \in \Sigma^* : \mu(x) \geq 2^{-H(\mu)}\}.$$

705 A standard “typical-set” argument shows that $\mu(S_{\text{high}}) \geq 1/2$ (or at least a fixed positive constant).

706 **Step 2: Missing S_{high} .** The probability that one sample $X \sim \mu$ does *not* land in S_{high} is at most
 707 $1 - \mu(S_{\text{high}}) \leq 1/2$. Hence the probability that *none* of N i.i.d. draws land in S_{high} is at most
 708 $(1/2)^N = 2^{-N}$. Thus any set of measure at least $1/2$ is missed with exponentially small probability.

709 **Step 3: Missing a smaller-mass region A .** If $\mu(A) = \epsilon \leq 1/2$, the probability that one draw misses
 710 A is $1 - \epsilon$, so the probability *all* N draws miss A is $(1 - \epsilon)^N \approx \exp(-N\epsilon)$ if $N\epsilon$ is not too large.

711 However, we need a *uniform* guarantee over *all* possible A of measure ϵ . By representing Σ^* as
 712 composed of at most $2^{H(\mu)}$ “atoms” each of probability at least $2^{-H(\mu)}$, the number of distinct
 713 subsets is at most $\exp(2^{H(\mu)} \ln 2)$. A union bound then modifies the exponent by about a factor of
 714 $1/2^{H(\mu)}$, so for large N one has

$$\Pr[\exists A : \mu(A) = \epsilon \text{ and all } N \text{ samples miss } A] \leq \exp\left(-\frac{N\epsilon}{2^{H(\mu)}}\right).$$

715 This is the desired exponential coverage bound.

716 **Step 4: Inverting the bound with the Lambert W -function.** We have

$$\Pr[\text{miss some set of mass } \epsilon] \leq \exp\left(-\frac{N\epsilon}{2^{H(\mu)}}\right).$$

717 We want to find ϵ such that this probability is itself at most ϵ :

$$\exp\left(-\frac{N\epsilon}{2^{H(\mu)}}\right) = \epsilon.$$

718 Rearrange as

$$\epsilon = \exp\left(-\frac{N\epsilon}{2^{H(\mu)}}\right) \iff \epsilon \exp\left(\frac{N\epsilon}{2^{H(\mu)}}\right) = 1.$$

719 Let $x = \frac{N\epsilon}{2^{H(\mu)}}$. Then $\epsilon = \frac{2^{H(\mu)}}{N}x$, and the above becomes

$$\frac{2^{H(\mu)}}{N} x \exp(x) = 1 \iff x e^x = \frac{N}{2^{H(\mu)}}.$$

720 By definition of the Lambert W -function, $x = W\left(\frac{N}{2^{H(\mu)}}\right)$. Substituting back, we get

$$\epsilon = \frac{2^{H(\mu)}}{N} W\left(\frac{N}{2^{H(\mu)}}\right).$$

721 Thus ϵ decreases to 0 as $N \rightarrow \infty$, roughly like

$$722 \frac{2^{H(\mu)}}{N} \ln\left(\frac{N}{2^{H(\mu)}}\right).$$

723 Hence, the probability of missing *any* set of mass at least ϵ is $\leq \epsilon$, with ϵ scaling on the order of
 724 $\frac{\ln(N)}{N}$. This completes the proof. \square

A.2 Temperature Sampling & Ablations

Sampling from an LLM at higher temperatures effectively “flattens” its probability distribution over next-token choices, increasing the entropy of the samples and thus encouraging exploration of lower-probability (more diverse) regions of the program space. Conversely, sampling at lower temperatures sharpens the distribution, concentrating probability mass on the model’s highest-confidence predictions and yielding lower-entropy (more conservative) samples. In other words, low-temperature sampling focuses on the most likely, canonical SMT-LIB programs (small effective support), while high-temperature sampling ventures into rarer, more varied corners of the output space (large effective support). If instead of a smoothly varying temperature schedule you simply draw many samples at fixed temperatures—say 0.5, 1.0, 1.5, and 2.0—you will still span low- to high-entropy regimes, but less systematically. You risk oversampling similar outputs at each temperature (especially near the extremes) and undersampling the intermediate entropy levels that lie between 0.5->1.0 and 1.5->2.0. A continuous schedule allocates exactly one sample per intermediate temperature, guaranteeing uniform coverage of entropy levels; fixed-temperature repetition may require substantially more draws to approximate that coverage, potentially leaving gaps in the distribution of generated programs.

Definition 1 (Gaussian Temperature Schedule). *To smoothly explore the distribution over SMT-LIB programs, we can define a temperature schedule for N samples as:*

$$\tau_i = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \exp\left(-\frac{(i - N/2)^2}{2\sigma^2}\right) \quad (1)$$

where $\tau_{\min} = 0.1$, $\tau_{\max} = 1.5$, and $\sigma = \frac{N}{5}$ controls the spread of the Gaussian.

We can also skew this gaussian towards lower temperatures.

Definition 2 (Exponential Temperature Schedule). *We can define a schedule that emphasizes sampling at lower temperatures using:*

$$\tau_i = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \exp(-\lambda \cdot i) \quad (2)$$

where $\lambda > 0$ controls the decay rate, and $i = 0, 1, \dots, N - 1$.

Comparison: From a purely coverage-guarantee standpoint (i.e. our goal of hitting every “significant” region of the SMT-LIB output distribution at least once), the Systematic uniform schedule remains the most theoretically justified. It uniformly samples every temperature exactly once, from low to high. Provably minimizes the worst-case “miss probability” by evenly covering the full entropy range. Gaussians concentrates samples near the middle temperature; fewer at extremes. It does provide smooth transitions; and avoids extreme high-entropy noise. However undersamples both very low-entropy (conservative) and very high-entropy (creative) regions resulting in weaker uniform coverage. The exponential decay schedule heavily biases toward low temperatures (low entropy), and therefore quickly focuses on high-confidence outputs. However, there is almost no exploration of rare programs; poor coverage of tail regions.

B Temperature-Varied SMT Generation and PCFG Analysis

To empirically investigate the influence of LLM sampling temperature on the characteristics of generated formal artifacts, we performed SMT-LIB v2 program generation across a defined temperature spectrum (e.g., $T_{\min} = 0.0$ to $T_{\max} = 2.0$). Distinct Probabilistic Context-Free Grammars (PCFGs) were induced from the SMT program ensembles parsed at each temperature point T_i , modeling the LLM’s syntactic and structural tendencies under each generative condition.

Our analysis of these per-temperature PCFGs revealed distinct and significant trends as sampling temperature was varied. Notably, the PCFG spectral radius generally trended upwards with increasing temperature. This intriguing behavior suggests that higher temperatures, while fostering diversity, may also enable the LLM to access and generate SMT structures with more pronounced or varied recursive complexity, perhaps by activating a broader range of complex production rules rather than simplifying structural choices. Consistent with expectations of increased diversity, grammar entropy and its associated perplexity also demonstrated an upward trend, quantifying the heightened uncertainty and the expanded set of effective choices exercised by the LLM at higher temperatures.

A particularly interesting observation was that the KL divergence from a uniform distribution also tended to increase with temperature. This implies that as the LLM explores a wider variety

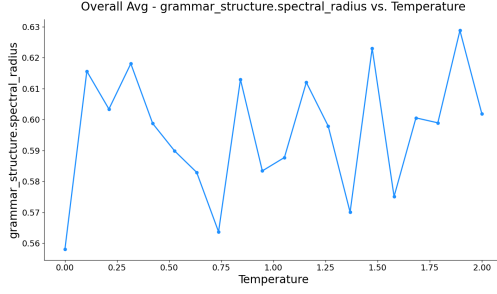


Figure 2: Spectral Radius VS Temperature

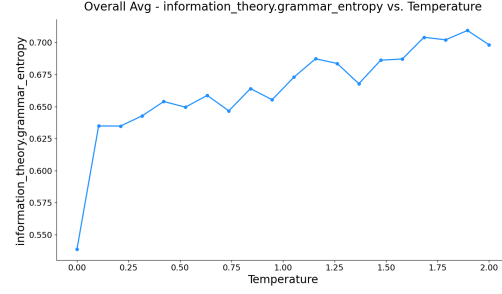


Figure 3: Grammar Entropy VS Temperature

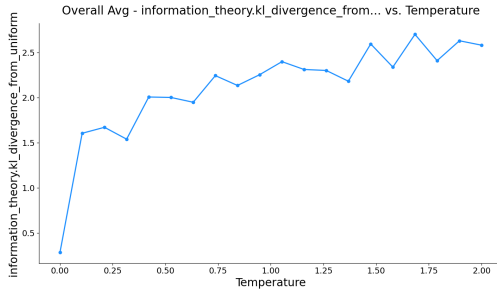


Figure 4: KLD vs Temp

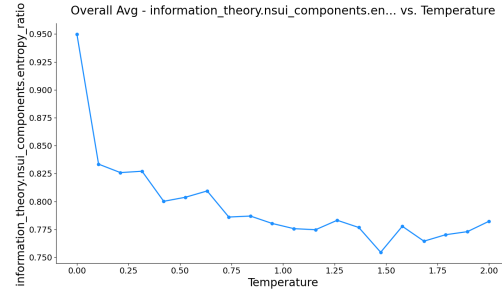


Figure 5: Entropy Ratio vs Temp

of production rules (evidenced by increased entropy), its choices within this expanded repertoire become, in a relative sense, more specific or structured, deviating further from a purely random uniform selection over the increasingly diverse set of utilized rules. Correspondingly, the entropy ratio generally decreased, which could occur if the maximum possible entropy (based on the growing set of observed rules and non-terminals at higher temperatures) increases at a faster rate than the actual grammar entropy. The composite metric NSUI showed fluctuating behavior without a clear monotonic direction, reflecting the complex interplay of its underlying components. The spectral factor, linked to the spectral radius, also exhibited a slight upward trend.

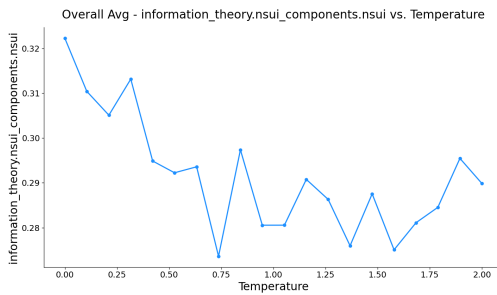


Figure 6: NSUI vs Temp

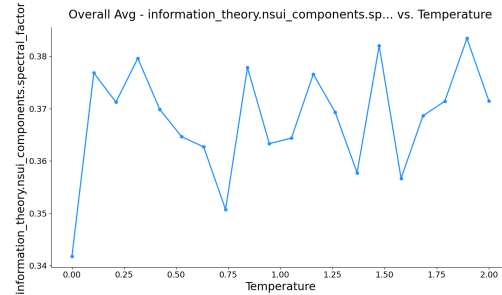


Figure 7: Spectral Factor vs Temp

Metrics related to the observed grammar structure, such as the average number of rules utilized per non-terminal, the maximum observed branching factor, and the average right-hand side (RHS) length of applied rules, all generally increased with temperature. This supports the notion that higher temperatures lead the LLM to explore and employ a more extensive and potentially more elaborate subset of the SMT-LIB grammar. Regarding the shape of the rule probability distributions, kurtosis consistently decreased, indicating that these distributions become flatter (less peaked) as temperature promotes more uniform rule selection among the actively used rules. Conversely, the skew of these

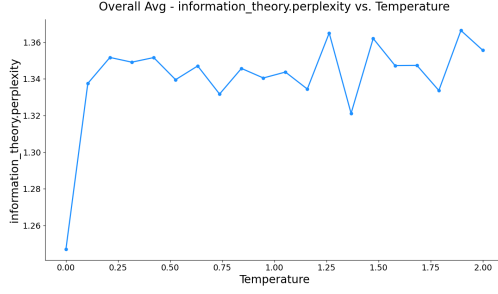


Figure 8: Perplexity vs Temp

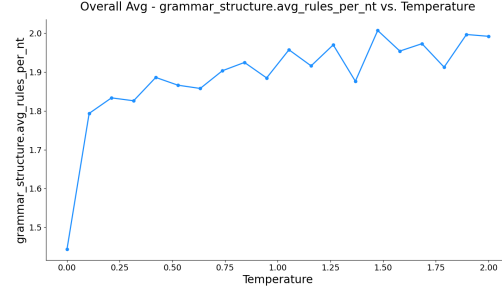


Figure 9: Average Rules per Non-terminal vs Temp

distributions tended to increase, suggesting a shift in the asymmetry of rule preferences as temperature changes.

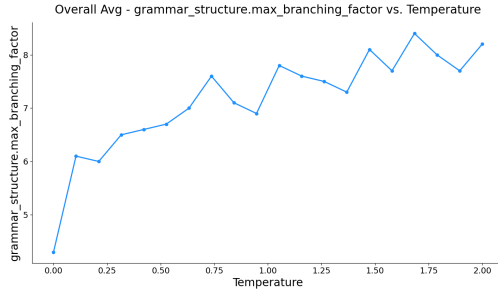


Figure 10: Max Branching Factor vs Temp

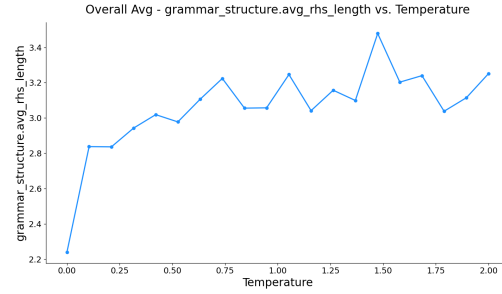


Figure 11: Average RHS Length vs Temp

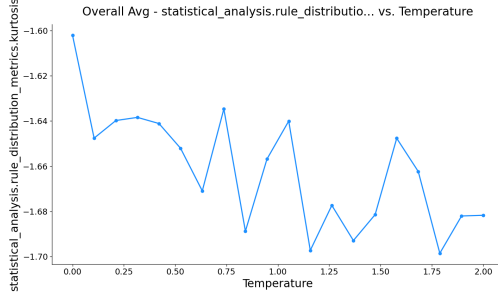


Figure 12: Kurtosis vs Temp

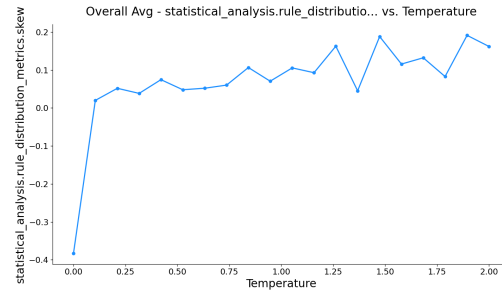


Figure 13: Skew vs Temp

These ablation studies are meaningful as they reveal a nuanced picture of the LLM’s generative process for formal languages. The trends suggest that increasing temperature doesn’t merely lead to random, uniform outputs, but rather allows the LLM to explore a richer, potentially more complex, and structurally diverse portion of the language space defined by G_{SMT} . This expansion, however, may also come with its own emergent structural specificities, as indicated by the KL divergence. These findings are crucial for understanding the coherence-diversity trade-off, for validating the sensitivity of PCFG-derived metrics, and for interpreting uncertainty scores, as the baseline characteristics of generated artifacts are systematically altered by temperature in complex ways. The observed responses underscore the value of empirical studies in characterizing LLM behavior for formal code generation.

C Detailed Results

C.1 Benchmarking Autoformalization

The performance benchmarks detailed in 5, 6, 7, 8, 9 were generated by evaluating five Large Language Models (o3-mini, DeepSeekR1 with Chain-of-Thought, DeepSeek-v3-04-21, Gemini Flash 2.0, and Gemini Flash 2.0 Lite) on four reasoning datasets. For each question, five samples were generated, and answers were derived either directly from the LLM’s textual output or by solving LLM-generated SMT-LIB programs using the Z3 solver. The “medium effort” designation for o3-mini indicates a specific prompting or iteration level for that model. In Table 5, the SMT approach for models like Deepseek v3 not only altered precision and recall but also resulted in substantial numbers of both False Positives (144) and True Positives (199), suggesting that while it attempted more proofs, a large fraction of these new attempts were erroneous. This contrasts with its text performance (42 FP, 174 TP). For the ProntoQA training set, with only true answers (Table 6), the SMT Precision of 1.0000 across all models is a direct consequence of the experimental design (no false statements to misclassify as true if TN is inherently 0); the variance in False Negatives (e.g., 270 for Deepseek v3 SMT) thus purely reflects the inability to successfully formalize and prove statements known to be true, a direct measure of formalization completeness for affirmatives.

Table 5: LLM Performance on StrategyQA (Text vs. SMT): SMT often boosts recall (e.g., Deepseek v3 from 0.81 to 0.91) but can reduce precision and overall accuracy for several models, highlighting model-dependent autoformalization success on knowledge-intensive tasks.

	StrategyQA															
	Text								SMT							
	Accuracy	Precision	Recall	F1	TP	TN	FP	FN	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
o3-mini (medium effort)	0.7828	0.8609	0.6047	0.7104	130	260	21	80	0.7980	0.8688	0.6347	0.7335	139	260	21	80
Deepseek v3 0324	0.8292	0.8055	0.8055	0.8055	174	234	42	42	0.6720	0.5801	0.9086	0.7081	199	137	144	20
DeepSeek R1	0.8580	0.8364	0.8402	0.8383	184	245	36	35	0.7760	0.7184	0.8037	0.7586	176	212	69	43
Gemini Flash 2.0	0.7188	0.6880	0.6570	0.6720	144	214	65	75	0.5360	0.4840	0.9269	0.6363	203	65	216	16
Gemini Flash 2.0 Lite	0.6760	0.6770	0.4970	0.5736	109	229	52	110	0.4500	0.4419	0.9726	0.6077	213	12	269	6

Table 6: LLM Performance on ProntoQA Train (True Statements Only; Text vs. SMT): SMT exposes significant failures in formalizing and proving known true statements for models like Deepseek v3 (Accuracy 0.45 vs. Text 1.00), indicating critical autoformalization recall deficiencies rather than precision issues (SMT Precision remains 1.00 for all).

	ProntoQA Train - ONLY TRUE Answers															
	Text								SMT							
	Accuracy	Precision	Recall	F1	TP	TN	FP	FN	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
o3-mini (medium effort)	1.0000	1.0000	1.0000	1.0000	499	0	0	0	0.9980	1.0000	0.9980	0.9889	499	0	0	1
Deepseek v3 0324	1.0000	1.0000	1.0000	1.0000	450	0	0	0	0.4501	1.0000	0.4501	0.6200	221	0	0	270
DeepSeek R1	0.9939	1.0000	0.9939	0.9969	489	0	0	3	0.7440	1.0000	0.7440	0.8532	372	0	0	128
Gemini Flash 2.0	0.9820	1.0000	0.9820	0.9900	491	0	0	9	0.9000	1.0000	0.9000	0.9470	450	0	0	50
Gemini Flash 2.0 Lite	0.9980	1.0000	0.9980	0.9980	499	0	0	1	0.9980	1.0000	0.9980	0.9989	499	0	0	1

The ProofWriter results Table 7 are notable. We advise the reader to ignore DeepSeek R1’s SMT performance, since it is based on a “Partial Run,” because of poor model ability to autoformalize, due to the overuse of thinking tokens, thereby causing an intractable timeline for converging to any solution and API call explosion. Here, o3-mini (medium effort) showcases a successful SMT application, where its accuracy improved to 0.9418 with a reduction in both False Positives (from 34 to 19) and False Negatives (from 21 to 10) compared to its text output. On the FOLIO dataset (Table 8), a common pattern observed in the SMT condition, beyond just low precision, was the significant reduction in True Negatives compared to the Text condition for several models (e.g., Deepseek v3 dropped from 34 TN via Text to 5 TN via SMT; Gemini Flash 2.0 from 37 TN to 0 TN). This suggests a systemic challenge in generating SMT formulas that correctly evaluate to unsatisfiable for statements that are indeed false within the FOLIO logical structure. Finally, the ProntoQA test set which includes both true and false statements (Table 9) revealed extreme model-specific behaviors under SMT. DeepSeek R1’s SMT output, for instance, correctly identified all 242 false statements (0 FP, 242 TN) but failed to correctly identify any of the 258 true statements (0 TP, 258 FN), indicating a systematic bias in its SMT generation towards unsatisfiability or an inability to complete proofs for satisfiable formulas in a mixed-distribution context, a stark contrast to its perfect text performance and its SMT performance on true-only statements.

835

Table 7: LLM Performance on ProofWriter (Text vs. SMT): SMT substantially improves models struggling with formal logic (e.g., Gemini Flash 2.0 Lite accuracy from 0.41 to 0.75), yet can degrade performance for models already strong in textual formal reasoning (e.g., DeepSeek R1 accuracy from 0.94 to 0.49), showcasing task-specific SMT utility.

	ProofWriter															
	Text								SMT							
	Accuracy	Precision	Recall	F1	TP	TN	FP	FN	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
o3-mini (medium effort)	0.8893	0.8697	0.9153	0.8919	227	215	34	21	0.9418	0.9261	0.9597	0.9426	238	231	19	10
Deepseek v3 0324	0.8057	0.8016	0.8225	0.8110	190	175	47	41	0.5800	0.6587	0.3320	0.4414	83	207	43	167
DeepSeek R1 (Partial Run)	0.9423	0.9597	0.9220	0.9400	143	151	6	12	0.4935	0.4750	0.1870	0.2685	29	125	32	126
Gemini Flash 2.0	0.4900	0.4960	0.5710	0.5300	140	106	142	105	0.6660	0.6844	0.6160	0.5313	154	106	71	96
Gemini Flash 2.0 Lite	0.4060	0.3609	0.2440	0.2911	61	142	108	189	0.7540	0.7275	0.8120	0.7674	203	174	76	47

836

Table 8: LLM Performance on FOLIO (Text vs. SMT): Textual reasoning largely outperforms SMT. For many models, SMT results in high recall but poor precision (e.g., Gemini Flash 2.0 SMT F1 0.72 vs Text 0.92) and a failure to identify false statements (e.g., 0 SMT True Negatives for Gemini Flash 2.0), indicating issues with formalizing negation or complex FOL conditions.

	Folio															
	Text								SMT							
	Accuracy	Precision	Recall	F1	TP	TN	FP	FN	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
o3-mini (medium effort)	0.9450	0.9682	0.9384	0.9531	61	43	2	4	0.5000	0.6890	0.2985	0.4166	20	36	9	47
Deepseek v3 0324	0.9333	0.9259	0.9615	0.9433	50	34	4	2	0.5961	0.6063	0.9193	0.7307	57	5	37	5
DeepSeek R1	0.9252	0.9670	0.9090	0.9374	60	39	2	6	0.5200	0.6363	0.5303	0.5785	35	21	20	31
Gemini Flash 2.0	0.9010	0.9275	0.9142	0.9200	64	37	5	6	0.5625	0.6000	0.9000	0.7200	63	0	42	7
Gemini Flash 2.0-lite	0.9017	0.904	0.9428	0.923	66	35	7	4	0.7321	0.7	1	0.8235	70	12	30	0

837

Table 9: LLM Performance on ProntoQA Test (True/False Mix; Text vs. SMT): SMT shows divergent outcomes: catastrophic failure for some (e.g., DeepSeek R1 SMT F1 0.00 vs. Text 1.00), yet significant improvement for others (Gemini Flash 2.0 Lite SMT Accuracy 0.78 vs. Text 0.56), highlighting inconsistent SMT reliability on mixed arithmetic statements.

	ProntoQA TEST - BOTH TRUE AND FALSE															
	Text								SMT							
	Accuracy	Precision	Recall	F1	TP	TN	FP	FN	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
o3-mini (medium effort)	1.0000	1.0000	1.0000	1.0000	258	240	0	0	1.0000	1.0000	1.0000	1.0000	258	242	0	0
Deepseek v3 0324	0.7200	0.7138	0.7635	0.7378	197	163	79	61	0.5140	0.5484	0.3295	0.4116	85	172	70	173
DeepSeek R1	1.0000	1.0000	1.0000	1.0000	253	242	0	0	0.4840	0.0000	0.0000	0.0000	0	242	0	258
Gemini Flash 2.0	0.7180	0.8232	0.5770	0.6780	149	210	32	109	0.4560	0.4753	0.5232	0.4981	135	93	149	123
Gemini Flash 2.0 Lite	0.5630	0.5811	0.5333	0.5562	136	144	98	119	0.7820	0.7210	0.9418	0.8168	243	148	94	15

C.2 Detailed Performance of Uncertainty Metrics for Ground Truth Prediction

This section provides a granular view of the performance of various Probabilistic Context-Free Grammar (PCFG) derived metrics, self-consistency measures, and ensemble methods in predicting the correctness of SMT-LIB outputs (with respect to ground truth) for specific LLM and dataset combinations. The metrics evaluated include AUROC (Area Under the Receiver Operating Characteristic Curve) for discrimination, ECE (Expected Calibration Error) and Brier score for calibration, and AURC (Area Under the Risk-Coverage Curve) along with optimal threshold (Opt.T), error rate at threshold (Err@T), and relative error reduction (RelErrRed) for selective prediction utility.

The UQ results for o3-mini on StrategyQA demonstrate moderate success in distinguishing correct SMT outputs from incorrect ones. While Grammar Entropy shows good individual discriminative power (AUROC 0.7448, AURC 0.1113), achieving a 13.88% relative error reduction by abstaining on just 5% of samples, many other standalone PCFG metrics exhibit weaker performance. The self-consistency metrics (Text and SMT) also perform well (AUROC 0.74), indicating that agreement between the LLM’s own reasoning modalities is a key signal. Notably, the Ensemble ML method achieves the highest AUROC (0.7850) and a significant relative error reduction (29.29% by abstaining on 10% of samples), underscoring the benefit of integrating diverse uncertainty signals through a learned model. The comparatively higher ECE for many metrics suggests that while discriminative, their raw scores may not always be well-calibrated probabilities.

Table 10: Uncertainty Quantification for o3-mini on StrategyQA: Ensemble ML (AUROC 0.7850) and Self-Consistency metrics (Text/SMT AUROC 0.74) outperform most individual PCFG metrics (Grammar Entropy AUROC 0.7448 being a strong contender). This suggests that for o3-mini on this knowledge-intensive task, learned combinations or behavioral consistency signals are more potent than isolated SMT structural properties for error detection.

Metric	StrategyQA - o3-mini						
	AUROC	ECE	Brier	AURC	Opt.T	Err@T	RelErrRed
Grammar Entropy	0.7448	0.3058	0.2340	0.1113	0.0500	0.1895	0.1388
Perplexity	0.5589	0.3107	0.2862	0.1811	0.2000	0.1750	0.2045
KL Divergence	0.6428	0.2485	0.2385	0.1471	0.3000	0.1429	0.3506
NSUI	0.6334	0.2436	0.2539	0.1250	0.1500	0.1882	0.1444
Renyi Ent (2)	0.5175	0.3303	0.2997	0.1977	0.3000	0.1857	0.1558
Renyi Ent (0.5)	0.5973	0.3398	0.3042	0.1634	0.2500	0.1600	0.2727
Max Ent	0.6649	0.3553	0.2935	0.1297	0.0500	0.1895	0.1388
Ent Ratio	0.5385	0.3283	0.3028	0.1834	0.3000	0.1857	0.1558
Spectral Factor	0.6334	0.2173	0.2364	0.1319	0.1500	0.1882	0.1444
Spectral Radius	0.6334	0.2892	0.2747	0.1319	0.1500	0.1882	0.1444
# Nonterminals	0.5111	0.3540	0.3188	0.2006	0.2000	0.2125	0.0341
# Rules	0.5548	0.2117	0.2385	0.1855	0.2000	0.1750	0.2045
Avg Rules / NT	0.5737	0.2400	0.2415	0.1752	0.2000	0.1625	0.2614
Avg RHS Len	0.5350	0.6141	0.5651	0.1979	0.0500	0.2105	0.0431
Max Branch Factor	0.5181	0.1500	0.1997	0.1990	0.1500	0.1882	0.1444
Rule Dist Mean	0.5740	0.3161	0.2836	0.1752	0.2000	0.1625	0.2614
Rule Dist StdDev	0.5291	0.3995	0.3517	0.1811	0.0500	0.2105	0.0431
Rule Dist Skew	0.5833	0.3178	0.2850	0.1689	0.1500	0.1765	0.1979
Rule Dist Kurtosis	0.5659	0.3948	0.3420	0.1785	0.0500	0.2000	0.0909
Self Consistency Text	0.7369	0.1604	0.1603	0.1081	0.1500	0.1529	0.3048
Self Consistency SMT	0.7416	0.1523	0.1609	0.1051	0.1500	0.1529	0.3048
Ensemble Average	0.7622	0.3724	0.2916	0.1103	0.1000	0.1556	0.2929
Ensemble Weighted	0.7657	0.1738	0.1617	0.1099	0.0500	0.1895	0.1388
Ensemble ML	0.7850	0.2090	0.1756	0.1013	0.1000	0.1556	0.2929
Ensemble Simple	0.6702	0.2055	0.2104	0.1410	0.2000	0.1500	0.3182

For DeepSeek-v3 on StrategyQA, UQ metrics show a somewhat different pattern compared to o3-mini. Ensemble ML again provides the best overall discrimination (AUROC 0.7709), achieving a relative error reduction of 8.96% by abstaining on 5% of the samples. Interestingly, several individual PCFG-derived metrics, such as Grammar Entropy (AUROC 0.7087), Max Entropy (AUROC 0.6851), and Spectral Factor/Radius (AUROC 0.6800), demonstrate better discriminative power than the self-consistency metrics (AUROCs 0.60-0.62). This suggests that for DeepSeek-v3 on this task, intrinsic structural characteristics of the generated SMT are more indicative of correctness than its consistency with textual outputs. While Ensemble Simple yields the highest relative error reduction (35.14%), this comes at the cost of a high abstention rate (Opt.T 0.50), indicating practical trade-offs in selective prediction.

Table 11: Uncertainty Quantification for DeepSeek-v3 on StrategyQA: Ensemble ML leads with an AUROC of 0.7709. Several individual PCFG-based metrics like Grammar Entropy (AUROC 0.7087) and Max Entropy (AUROC 0.6851) show reasonable efficacy, outperforming self-consistency measures for this model.

Metric	AUROC	ECE	Brier	AURC	Opt.T	Err@T	RelErrRed
Grammar Entropy	0.7087	0.1575	0.2302	0.2097	0.05	0.3474	0.0612
Perplexity	0.6122	0.1601	0.2641	0.2497	0.5	0.28	0.2432
KL Divergence	0.5723	0.1393	0.2322	0.2878	0.05	0.3579	0.0327
NSUI	0.5997	0.0781	0.2191	0.2672	0.1	0.3222	0.1291
Renyi Ent (2)	0.6195	0.1622	0.2679	0.2429	0.45	0.2727	0.2629
Renyi Ent (0.5)	0.6126	0.1623	0.2626	0.2517	0.5	0.28	0.2432
Max Ent	0.6851	0.0473	0.2099	0.2271	0.1	0.3222	0.1291
Ent Ratio	0.5311	0.1336	0.2538	0.3306	0.05	0.3684	0.0043
Spectral Factor	0.68	0.0992	0.2236	0.2365	0.05	0.3474	0.0612
Spectral Radius	0.68	0.0686	0.2148	0.2365	0.05	0.3474	0.0612
# Nonterminals	0.6115	0.1329	0.2469	0.2547	0.45	0.2909	0.2138
# Rules	0.6197	0.1109	0.2252	0.2583	0.15	0.3176	0.1415
Avg Rules / NT	0.6021	0.0902	0.226	0.2616	0.05	0.3579	0.0327
Avg RHS Len	0.5122	0.1753	0.2712	0.3279	0.1	0.3444	0.0691
Max Branch Factor	0.618	0.145	0.227	0.2688	0.05	0.3474	0.0612
Rule Dist Mean	0.6021	0.1811	0.2534	0.2616	0.05	0.3579	0.0327
Rule Dist StdDev	0.5281	0.1251	0.2573	0.3116	0.5	0.32	0.1351
Rule Dist Skew	0.6036	0.1431	0.2489	0.261	0.1	0.3444	0.0691
Rule Dist Kurtosis	0.5787	0.172	0.26	0.2754	0.05	0.3579	0.0327
Self Consistency Text	0.6017	0.2874	0.3048	0.2882	0.1	0.3444	0.0691
Self Consistency SMT	0.6203	0.2318	0.2745	0.2513	0.1	0.3444	0.0691
Ensemble Average	0.6795	0.1214	0.2077	0.2182	0.1	0.3222	0.1291
Ensemble Weighted	0.7211	0.1257	0.2135	0.1989	0.05	0.3474	0.0612
Ensemble ML	0.7709	0.0877	0.1968	0.1847	0.05	0.3368	0.0896
Ensemble Simple	0.6401	0.1763	0.2514	0.2483	0.5	0.24	0.3514

866 The UQ performance for o3-mini on ProofWriter is remarkably high, demonstrating the strong
867 potential of PCFG-based metrics in formal reasoning contexts. Numerous individual metrics, in-
868 cluding Grammar Entropy, Perplexity (AUROC 0.9194), Renyi Entropy (0.5) (AUROC 0.9301),
869 Average Rules / NT (AUROC 0.9301), and various rule distribution statistics, achieve AUROC
870 scores exceeding 0.90. More impressively, their AURC values are exceptionally low (e.g., 0.0008 for
871 Grammar Entropy), translating to a 100% relative error reduction by abstaining on a small fraction of
872 samples (e.g., 10%). This strongly supports the hypothesis that syntactic irregularities in generated
873 formal artifacts are highly indicative of underlying semantic errors when the task aligns well with the
874 formal language. Ensemble methods elevate this performance to near-perfection (Ensemble Average
875 AUROC 0.9949). Despite the excellent discrimination, some metrics show high ECE values (e.g.,
876 Grammar Entropy ECE 0.4419), suggesting that while they can effectively rank outputs by correctness
877 likelihood, their raw scores may not be perfectly calibrated across the entire probability spectrum.

878 The UQ results for Gemini 2.0 Flash Lite on ProofWriter present a mixed picture, contrasting with
879 o3-mini’s strong performance on the same task. Many individual PCFG-derived metrics demonstrate
880 weak discriminative ability, with AUROC scores often between 0.50 and 0.59 (e.g., Grammar Entropy
881 at 0.5380, Spectral Radius at 0.5011). However, SMT Self Consistency stands out as a significantly
882 stronger individual performer with an AUROC of 0.7364. Ensemble methods, particularly Ensemble
883 ML, achieve the best overall performance (AUROC 0.7631, AURC 0.0823), leading to a 14.61%
884 relative error reduction when abstaining on 5% of the samples. This suggests that for Gemini 2.0
885 Flash Lite on ProofWriter, the structural variations in its SMT outputs are less consistently tied to
886 semantic correctness compared to o3-mini. Instead, behavioral consistency (specifically, how its SMT
887 outputs align with each other across multiple generations) and learned patterns across a combination
888 of (often individually weaker) signals provide more reliable error detection.

Table 12: Uncertainty Quantification for o3-mini on ProofWriter: PCFG-derived metrics achieve exceptional discriminative power (e.g., Grammar Entropy AUROC 0.9301, AURC 0.0008), enabling near-perfect error detection with minimal abstention (100% RelErrRed at Opt.T 0.10). Ensemble methods (e.g., Ensemble Average AUROC 0.9949) further refine this, confirming that SMT structural properties are extremely strong predictors of correctness for o3-mini on this formal reasoning task.

Metric	ProofWriter						RelErrRed
	AUROC	ECE	Brier	AURC	Opt.T	Err@T	
Grammar Entropy	0.9301	0.4419	0.25	0.0008	0.1000	0.0000	1.0000
Perplexity	0.9194	0.5358	0.3515	0.0008	0.1000	0.0000	1.0000
KL Divergence	0.5108	0.5167	0.326	0.0074	0.0000	0.0106	0.0000
NSUI	0.5645	0.571	0.3843	0.0084	0.0000	0.0106	0.0000
Renyi Ent (2)	0.8871	0.5405	0.3598	0.0013	0.1500	0.0000	1.0000
Renyi Ent (0.5)	0.9301	0.4724	0.2879	0.0008	0.1000	0.0000	1.0000
Max Ent	0.9086	0.7198	0.555	0.0013	0.1000	0.0000	1.0000
Ent Ratio	0.586	0.5714	0.3764	0.0055	0.4500	0.0000	1.0000
Spectral Factor	0.7473	0.3458	0.2247	0.0032	0.3000	0.0000	1.0000
Spectral Radius	0.7473	0.3545	0.2305	0.0032	0.3000	0.0000	1.0000
# Nonterminals	0.8011	0.4267	0.2397	0.0019	0.2000	0.0000	1.0000
# Rules	0.5108	0.4186	0.2201	0.0084	0.0000	0.0106	0.0000
Avg Rules / NT	0.9301	0.5393	0.3499	0.0008	0.1000	0.0000	1.0000
Avg RHS Len	0.8011	0.6535	0.5086	0.0026	0.2500	0.0000	1.0000
Max Branch Factor	0.5914	0.2979	0.1377	0.0055	0.4500	0.0000	1.0000
Rule Dist Mean	0.9301	0.4945	0.3034	0.0008	0.1000	0.0000	1.0000
Rule Dist StdDev	0.5108	0.5555	0.3723	0.0074	0.5000	0.0000	1.0000
Rule Dist Skew	0.9301	0.4923	0.2987	0.0008	0.1000	0.0000	1.0000
Rule Dist Kurtosis	0.586	0.5107	0.3115	0.0055	0.4500	0.0000	1.0000
Self Consistency Text	0.899	0.0423	0.028	0.002	0.0500	0.0105	0.4684
Self Consistency SMT	0.7121	0.8501	0.7764	0.0025	0.1500	0.0000	1.0000
Ensemble Average	0.9949	0.3356	0.1414	0.0005	0.0500	0.0000	1.0000
Ensemble Weighted	0.9785	0.4612	0.2566	0.0003	0.0500	0.0000	1.0000
Ensemble ML	0.9892	0.0572	0.028	0.0003	0.0500	0.0000	1.0000
Ensemble Simple	0.9355	0.4419	0.2582	0.0008	0.1000	0.0000	1.0000

Table 13: Uncertainty Quantification for Gemini 2.0 Flash Lite on ProofWriter: Performance is moderate; SMT Self Consistency (AUROC 0.7364) and Ensemble ML (AUROC 0.7631) are the strongest UQ signals. Most individual PCFG structural metrics show weak discriminative power (many AUROCs 0.50-0.59), indicating that for this model on ProofWriter, behavioral consistency (SMT-based) and learned combinations are more indicative of correctness than raw SMT syntactic properties alone.

Metric	ProofWriter						RelErrRed
	AUROC	ECE	Brier	AURC	Opt.T	Err@T	
Grammar Entropy	0.5380	0.3185	0.2869	0.1405	0.4500	0.1667	0.1081
Perplexity	0.5934	0.3888	0.3182	0.1267	0.1000	0.1742	0.0680
KL Divergence	0.5164	0.3080	0.2797	0.1573	0.0000	0.1869	0.0000
NSUI	0.5243	0.3186	0.2642	0.1514	0.1500	0.1845	0.0125
Renyi Ent (2)	0.5996	0.4102	0.3368	0.1285	0.1000	0.1742	0.0680
Renyi Ent (0.5)	0.5933	0.4401	0.3581	0.1258	0.1000	0.1742	0.0680
Max Ent	0.5417	0.3503	0.3045	0.1420	0.5000	0.1717	0.0811
Ent Ratio	0.5177	0.3943	0.3426	0.1548	0.2000	0.1835	0.0178
Spectral Factor	0.5011	0.5048	0.4157	0.1578	0.0000	0.1869	0.0000
Spectral Radius	0.5011	0.3930	0.3172	0.1578	0.0000	0.1869	0.0000
# Nonterminals	0.5167	0.4838	0.4215	0.1672	0.1000	0.1854	0.0079
# Rules	0.5549	0.2422	0.2315	0.1370	0.4000	0.1610	0.1383
Avg Rules / NT	0.5840	0.2790	0.2656	0.1301	0.3000	0.1377	0.2632
Avg RHS Len	0.5631	0.3413	0.2906	0.1480	0.1000	0.1685	0.0981
Max Branch Factor	0.5745	0.2189	0.2293	0.1355	0.3000	0.1522	0.1857
Rule Dist Mean	0.5838	0.4368	0.3713	0.1301	0.3000	0.1377	0.2632
Rule Dist StdDev	0.5144	0.4474	0.3795	0.1559	0.3500	0.1797	0.0384
Rule Dist Skew	0.5844	0.4511	0.3779	0.1313	0.3000	0.1522	0.1857
Rule Dist Kurtosis	0.5044	0.1437	0.1913	0.1726	0.4000	0.1610	0.1383
Self Consistency Text	0.5525	0.2283	0.2419	0.1376	0.4500	0.1545	0.1866
Self Consistency SMT	0.7364	0.3535	0.2751	0.1031	0.2000	0.1062	0.4408
Ensemble Average	0.6140	0.3922	0.3192	0.1240	0.1000	0.1722	0.0936
Ensemble Weighted	0.7235	0.3327	0.2539	0.1035	0.1000	0.1404	0.2484
Ensemble ML	0.7631	0.2897	0.2229	0.0823	0.0500	0.1596	0.1461
Ensemble Simple	0.6476	0.3867	0.3039	0.1071	0.2000	0.1519	0.1871

C.3 Detailed Performance of SMT-Based Uncertainty Metrics for Text-Answer Prediction

This section evaluates the efficacy of uncertainty quantification (UQ) metrics derived from SMT-LIB generations in predicting the correctness of the SMT results with the corresponding textual answers. The goal is to identify when the formalization (SMT output) aligns or diverges from the model’s natural language reasoning output (textual answer). On StrategyQA, o3 mini had 100% agreement between SMT and text answers, so UQ analysis for SMT-Text consistency prediction was not applicable for that specific model-dataset pair as there were no disagreements to predict. Results for other cases are detailed below.

For DeepSeek R1 on StrategyQA, we assesses how well metrics derived from its SMT generations can predict alignment with its textual answers. The results are strong: ensemble methods integrating these SMT features, such as Ensemble Weighted (AUROC 0.8494) and Ensemble Average (AUROC 0.8183), are highly effective. Notably, Text Self Consistency (AUROC 0.8245) is a top individual performer, suggesting that instability in textual outputs often correlates with SMT-Text divergence. Among metrics purely derived from SMT structure, Grammar Entropy (AUROC 0.7609) is noteworthy, achieving a 100% relative error reduction in identifying SMT-Text disagreements if one abstains on 45% of cases. This performance in predicting SMT-Text consistency is robust and highlights that both SMT structural integrity and textual stability are key indicators. The AURC values are generally very low for top performers (e.g., 0.0155 for Ensemble Weighted), indicating high utility in selectively flagging potential cross-modal disagreements.

Table 14: UQ for SMT-Text Consistency (DeepSeek R1, StrategyQA): SMT-derived metrics, especially ensembles (Ensemble Weighted AUROC 0.8494), effectively predict SMT-Text answer agreement. Text Self Consistency (AUROC 0.8245) is a strong predictor, while SMT-derived Grammar Entropy (AUROC 0.7609) also shows good utility, enabling high error reduction (100% RelErrRed at Opt.T 0.45) in identifying SMT-Text divergences.

Metric	AUROC	ECE	Brier	AURC	Opt.T	Err@T	RelErrRed
Grammar Entropy	0.7609	0.4617	0.2968	0.0216	0.4500	0.0000	1.0000
Perplexity	0.6211	0.3789	0.2640	0.0361	0.5000	0.0204	0.7114
KL Divergence	0.5776	0.4408	0.2855	0.0443	0.3000	0.0580	0.1801
NSUI	0.5963	0.2529	0.1433	0.0462	0.1000	0.0562	0.2055
Renyi Ent (2)	0.6242	0.3980	0.2746	0.0378	0.5000	0.0204	0.7114
Renyi Ent (0.5)	0.6149	0.3942	0.2755	0.0376	0.5000	0.0408	0.4227
Max Ent	0.7174	0.3994	0.2341	0.0262	0.0500	0.0638	0.0973
Ent Ratio	0.5217	0.5047	0.3529	0.0543	0.3000	0.0580	0.1801
Spectral Factor	0.5481	0.1533	0.0927	0.0640	0.1500	0.0476	0.3265
Spectral Radius	0.5481	0.2067	0.1166	0.0640	0.1500	0.0476	0.3265
# Nonterminals	0.5264	0.5679	0.4237	0.0557	0.5000	0.0408	0.4227
# Rules	0.6087	0.3083	0.1839	0.0396	0.4000	0.0508	0.2809
Avg Rules / NT	0.7034	0.4068	0.2532	0.0273	0.0500	0.0638	0.0973
Avg RHS Len	0.5637	0.2491	0.1566	0.0544	0.1000	0.0562	0.2055
Max Branch Factor	0.6801	0.3325	0.2042	0.0324	0.0500	0.0638	0.0973
Rule Dist Mean	0.7034	0.5352	0.3733	0.0273	0.0500	0.0638	0.0973
Rule Dist StdDev	0.6056	0.6755	0.5361	0.0413	0.2000	0.0506	0.2839
Rule Dist Skew	0.6848	0.4800	0.3260	0.0309	0.0500	0.0638	0.0973
Rule Dist Kurtosis	0.5311	0.2521	0.1588	0.0534	0.1000	0.0674	0.0465
Self Consistency Text	0.8245	0.1002	0.0751	0.0155	0.0500	0.0319	0.5486
Self Consistency SMT	0.7570	0.2062	0.1373	0.0268	0.0500	0.0532	0.2477
Ensemble Average	0.8183	0.4695	0.3058	0.0180	0.2500	0.0135	0.8089
Ensemble Weighted	0.8494	0.3256	0.1711	0.0155	0.0500	0.0319	0.5486
Ensemble ML	0.8245	0.3084	0.2003	0.0170	0.2000	0.0380	0.4629
Ensemble Simple	0.6957	0.4363	0.2748	0.0316	0.1000	0.0562	0.2055

When predicting SMT-Text consistency for DeepSeek v3 on StrategyQA, UQ metrics based on SMT generations prove highly effective. The Ensemble ML approach, which learns from various SMT-derived features, achieves an impressive AUROC of 0.8517 and offers a 55.56% relative error reduction in spotting SMT-Text disagreements when abstaining on 25% of samples. Good individual predictors include SMT-derived Grammar Entropy (AUROC 0.7354) and SMT Self Consistency (AUROC 0.7116). This demonstrates that for DeepSeek v3, deviations from typical SMT structure (signaled by grammar entropy) or inconsistencies in the SMT generation process itself are strong indicators that the SMT output might not align with the model’s textual answer. The low AURC (0.0573) for Ensemble ML highlights its practical utility. This task of predicting internal consistency (SMT-Text) shows strong signals, comparable to or even clearer (e.g. for Ensemble ML) than predicting SMT-Ground Truth correctness for this model on the same dataset.

Table 15: UQ for SMT-Text Consistency (DeepSeek v3, StrategyQA): Ensemble ML using SMT-derived features shows excellent performance (AUROC 0.8517) in predicting SMT-Text agreement, with a 55.56% relative error reduction. SMT-derived Grammar Entropy (AUROC 0.7354) and SMT Self Consistency (AUROC 0.7116) also serve as solid individual predictors, indicating that atypical SMT structures and generation instability can flag potential SMT-Text divergences.

Metric	AUROC	ECE	Brier	AURC	Opt.T	Err@T	RelErrRed
Grammar Entropy	0.7354	0.3058	0.2551	0.097	0.0500	0.1789	0.1479
Perplexity	0.6721	0.3282	0.2718	0.1205	0.3000	0.1143	0.4558
KL Divergence	0.5335	0.1195	0.178	0.1633	0.0500	0.2000	0.0476
NSUI	0.5973	0.1768	0.1919	0.1411	0.0500	0.1895	0.0977
Renyi Ent (2)	0.6799	0.3403	0.2808	0.116	0.3000	0.1000	0.5238
Renyi Ent (0.5)	0.6667	0.3333	0.2751	0.1223	0.3500	0.1077	0.4872
Max Ent	0.6353	0.1309	0.1738	0.1247	0.1000	0.1889	0.1005
Ent Ratio	0.6154	0.4091	0.3307	0.1446	0.5000	0.1000	0.5238
Spectral Factor	0.7034	0.1254	0.158	0.1206	0.1000	0.1667	0.2063
Spectral Radius	0.7034	0.1896	0.1752	0.1206	0.1000	0.1667	0.2063
# Nonterminals	0.5549	0.2271	0.2455	0.148	0.0500	0.2000	0.0476
# Rules	0.5675	0.2025	0.2077	0.1485	0.1000	0.1778	0.1534
Avg Rules / NT	0.6034	0.1393	0.1854	0.1399	0.0500	0.1895	0.0977
Avg RHS Len	0.5208	0.1254	0.1818	0.1972	0.0500	0.2000	0.0476
Max Branch Factor	0.5937	0.1563	0.192	0.1457	0.1000	0.1889	0.1005
Rule Dist Mean	0.6034	0.321	0.2742	0.1399	0.0500	0.1895	0.0977
Rule Dist StdDev	0.6281	0.2226	0.2221	0.1445	0.3000	0.1571	0.2517
Rule Dist Skew	0.5986	0.3057	0.2619	0.1406	0.0500	0.1895	0.0977
Rule Dist Kurtosis	0.66	0.0731	0.1576	0.1256	0.0500	0.1895	0.0977
Self Consistency Text	0.5778	0.1874	0.2008	0.1754	0.1500	0.1765	0.1597
Self Consistency SMT	0.7116	0.1662	0.1909	0.1054	0.1000	0.1778	0.1534
Ensemble Average	0.7064	0.1876	0.1831	0.0983	0.1000	0.1667	0.2063
Ensemble Weighted	0.7709	0.234	0.1971	0.0798	0.0500	0.1895	0.0977
Ensemble ML	0.8517	0.1861	0.1703	0.0573	0.2500	0.0933	0.5556
Ensemble Simple	0.6727	0.3021	0.2567	0.1119	0.1500	0.1765	0.1597

For Gemini Flash 2.0 Lite on StrategyQA, the task of predicting SMT-Text consistency reveals a standout individual metric: Rule Distribution Kurtosis from the SMT generations achieves a very high AUROC of 0.8695. This is a particularly interesting finding, as it suggests that the "tailedness" or outlier presence in the distribution of PCFG rules used during SMT generation is a very strong signal of whether the SMT output will align with the textual answer for this model. This metric's performance surpasses many other individual PCFG metrics (e.g., Grammar Entropy AUROC 0.6622, Perplexity AUROC 0.7212). Ensemble methods, like Ensemble Weighted (AUROC 0.8070) and Ensemble Average (AUROC 0.7927), provide robust overall performance, leveraging combinations of signals. The strong performance of kurtosis aligns with our discussion about "syntactic fingerprints" and how atypical SMT patterns (like bimodal distributions captured by kurtosis) can signal reasoning issues or misalignments. The AURC for Kurtosis (0.4448) suggests that while discriminative, its practical utility in terms of risk reduction might require careful thresholding, achieving a 30.56% error reduction at a 50% abstention rate.

The results for o3-mini on ProofWriter for predicting SMT-Text consistency are exceptional. Ensemble ML and Ensemble Weighted methods achieve perfect AUROC scores of 1.0000, signifying an ability to flawlessly distinguish SMT outputs that align with textual answers from those that diverge. This allows for a 100% relative error reduction with a very low 5% abstention rate. Beyond ensembles, many individual PCFG metrics derived from the SMT generations show extremely high predictive capabilities. For instance, Ent Ratio (AUROC 0.9677), Rule Dist Kurtosis (AUROC 0.9462), Max Ent (AUROC 0.9355), and NSUI (AUROC 0.9355) are all remarkably strong predictors, each achieving 100% relative error reduction at their respective optimal thresholds. This indicates that for o3-mini, particularly on a formal reasoning task like ProofWriter, the structural and probabilistic characteristics of its SMT generations are almost perfectly indicative of whether its formal and textual reasoning pathways are aligned. The exceptionally low AURC values (e.g., 0.0003 for Ensemble ML) further emphasize the practical certainty offered by these UQ measures in this context. This level of predictability for SMT-Text consistency is even more pronounced than some of the SMT-Ground Truth prediction results for this model, demonstrating the power of SMT features for diagnosing internal reasoning coherence.

Table 16: UQ for SMT-Text Consistency (Gemini Flash 2.0 Lite, StrategyQA): Rule Distribution Kurtosis (AUROC 0.8695) from SMT generations is an exceptionally strong individual predictor of SMT-Text agreement, significantly outperforming other PCFG metrics. Ensemble methods (e.g., Ensemble Weighted AUROC 0.8070) also perform well. This highlights a specific SMT structural feature as a key indicator of cross-modal alignment for this model.

Metric	AUROC	ECE	Brier	AURC	Opt.T	Err@T	RelErrRed
Grammar Entropy	0.6622	0.1277	0.2095	0.5689	0.1000	0.6889	0.0432
Perplexity	0.7212	0.3255	0.2849	0.5273	0.0500	0.7053	0.0205
KL Divergence	0.5486	0.2387	0.263	0.6341	0.1000	0.7000	0.0278
NSUI	0.6741	0.1195	0.1624	0.5221	0.5000	0.6600	0.0833
Renyi Ent (2)	0.7624	0.3192	0.2624	0.5037	0.1000	0.6889	0.0432
Renyi Ent (0.5)	0.6994	0.2766	0.2619	0.5438	0.0500	0.7053	0.0205
Max Ent	0.5982	0.3401	0.3083	0.59	0.3000	0.6714	0.0675
Ent Ratio	0.5511	0.1641	0.2511	0.6346	0.2500	0.6667	0.0741
Spectral Factor	0.6538	0.0943	0.1677	0.5316	0.5000	0.6600	0.0833
Spectral Radius	0.6538	0.1648	0.1703	0.5316	0.5000	0.6600	0.0833
# Nonterminals	0.5166	0.418	0.3958	0.6509	0.3500	0.6769	0.0598
# Rules	0.5749	0.4256	0.3818	0.6252	0.1000	0.6889	0.0432
Avg Rules / NT	0.6565	0.2913	0.2761	0.5923	0.2500	0.6400	0.1111
Avg RHS Len	0.6014	0.4855	0.442	0.6098	0.0500	0.7053	0.0205
Max Branch Factor	0.5818	0.5167	0.4747	0.6313	0.0500	0.7053	0.0205
Rule Dist Mean	0.6565	0.2197	0.228	0.5923	0.2500	0.6400	0.1111
Rule Dist StdDev	0.7183	0.3136	0.2807	0.5455	0.1500	0.6706	0.0686
Rule Dist Skew	0.6796	0.1783	0.2199	0.572	0.2500	0.6400	0.1111
Rule Dist Kurtosis	0.8695	0.3187	0.2412	0.4448	0.5000	0.5000	0.3056
Self Consistency Text	0.535	0.2788	0.2616	0.6064	0.3500	0.6462	0.1026
Self Consistency SMT	0.7505	0.538	0.4357	0.4822	0.5000	0.5600	0.2222
Ensemble Average	0.7927	0.1531	0.1702	0.4848	0.4000	0.5833	0.1898
Ensemble Weighted	0.807	0.1673	0.1632	0.4718	0.3000	0.6143	0.1468
Ensemble ML	0.7946	0.1308	0.1592	0.4784	0.5000	0.5400	0.2500
Ensemble Simple	0.7584	0.0796	0.1568	0.4929	0.1500	0.6824	0.0523

Table 17: UQ for SMT-Text Consistency (o3-mini, ProofWriter): SMT-derived UQ metrics demonstrate outstanding performance, with Ensemble ML and Ensemble Weighted achieving perfect AUROC (1.0000) in predicting SMT-Text agreement. Numerous individual PCFG metrics, such as Ent Ratio (AUROC 0.9677) and Rule Dist Kurtosis (AUROC 0.9462), are also exceptionally effective, enabling complete identification of SMT-Text inconsistencies with minimal abstention. This underscores a very strong link between SMT formalization properties and cross-modal reasoning alignment for o3-mini on this formal task.

Metric	AUROC	ECE	Brier	AURC	Opt.T	Err@T	RelErrRed
Grammar Entropy	0.8602	0.4419	0.2542	0.0013	0.15	0.00	1.00
Perplexity	0.9032	0.4429	0.2616	0.0013	0.10	0.00	1.00
KL Divergence	0.6667	0.5167	0.3238	0.0039	0.35	0.00	1.00
NSUI	0.9355	0.4077	0.2172	0.0008	0.10	0.00	1.00
Renyi Ent (2)	0.9032	0.4382	0.2580	0.0013	0.10	0.00	1.00
Renyi Ent (0.5)	0.8925	0.5063	0.3246	0.0013	0.15	0.00	1.00
Max Ent	0.9355	0.2589	0.1029	0.0008	0.10	0.00	1.00
Ent Ratio	0.9677	0.4074	0.2082	0.0003	0.05	0.00	1.00
Spectral Factor	0.5269	0.6329	0.5061	0.0084	0.00	0.01	0.00
Spectral Radius	0.5269	0.6243	0.4953	0.0084	0.00	0.01	0.00
# Nonterminals	0.6505	0.5520	0.3650	0.0047	0.40	0.00	1.00
# Rules	0.5108	0.4186	0.2201	0.0064	0.50	0.00	1.00
Avg Rules / NT	0.8011	0.4394	0.2554	0.0026	0.25	0.00	1.00
Avg RHS Len	0.5054	0.6535	0.5116	0.0074	0.50	0.00	1.00
Max Branch Factor	0.7419	0.6809	0.5100	0.0032	0.30	0.00	1.00
Rule Dist Mean	0.8011	0.4842	0.2971	0.0026	0.25	0.00	1.00
Rule Dist StdDev	0.8710	0.4232	0.2392	0.0013	0.15	0.00	1.00
Rule Dist Skew	0.8172	0.4864	0.2964	0.0019	0.20	0.00	1.00
Rule Dist Kurtosis	0.9462	0.5107	0.3056	0.0008	0.10	0.00	1.00
Self Consistency Text	0.7050	0.9352	0.9023	0.0030	0.30	0.00	1.00
Self Consistency SMT	0.7100	0.8600	0.7863	0.0030	0.30	0.00	1.00
Ensemble Average	0.9300	0.3560	0.1741	0.0007	0.10	0.00	1.00
Ensemble Weighted	1.0000	0.4231	0.2375	0.0003	0.05	0.00	1.00
Ensemble ML	1.0000	0.0496	0.0199	0.0003	0.05	0.00	1.00
Ensemble Simple	0.6667	0.4419	0.2661	0.0039	0.35	0.00	1.00

947 D Supplementary Experimental Details

948 The comprehensive PCFG analysis underpinning our uncertainty quantification was conducted on a
 949 focused set of benchmarks. Specifically, for 100 questions each from the StrategyQA, ProofWriter,
 950 and ProntoQA datasets, a corpus of $N_{SMT} = 100$ SMT-LIB v2 program samples per question was
 951 generated. The FOLIO dataset was excluded from this detailed PCFG study due to challenges in
 952 obtaining consistently robust SMT formalizations from the evaluated LLMs. Each SMT program
 953 within these corpora was parsed using an ANTLR-based parser to extract its constituent production
 954 rules. For the generation of these primary SMT samples used in uncertainty quantification (distinct
 955 from the temperature ablation study), LLM sampling temperature was maintained at its default setting
 956 to promote more deterministic outputs, with up to 10 generation attempts per SMT sample to ensure
 957 corpus completeness.

958 For each of the selected questions, a unique PCFG was induced from its corresponding 100 SMT
 959 samples. Rule probabilities within these per-question PCFGs were estimated via Maximum Likelihood
 960 Estimation (MLE), incorporating Lidstone smoothing (specifically, Laplace smoothing with $\beta_s = 1$)
 961 to manage unseen production rules. Beyond the metrics detailed in the main methodology, specific
 962 configurations included the computation of Rényi entropy for orders $\alpha = 0.5$ and $\alpha = 2.0$ (Collision
 963 Entropy).

964 The evaluation framework for the derived uncertainty metrics incorporated specific settings. Expected
 965 Calibration Error (ECE) was calculated using 10 discretization bins for confidence scores. In the
 966 analysis of selective prediction utility (error vs. abstention), optimal abstention thresholds were
 967 determined by targeting maximum relative error reduction while considering abstention levels up
 968 to a practical maximum of 50%. For our Ensemble ML predictor, a Logistic Regression model was
 969 employed, configured with balanced class weights and trained for up to 10,000 iterations on scaled
 970 features derived from the suite of PCFG uncertainty metrics.

971 E SMT Error Ratios vs Text Error Ratios

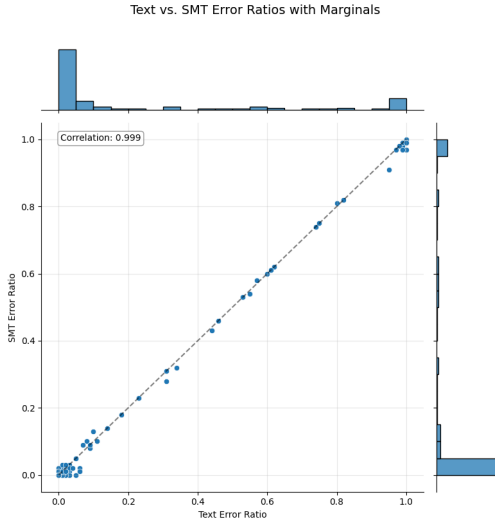


Figure 14: SMT vs. Text Error Ratio Analysis for o3-mini: Illustrates well-calibrated SMT generation, indicated by a strong correlation between SMT and Text error patterns.

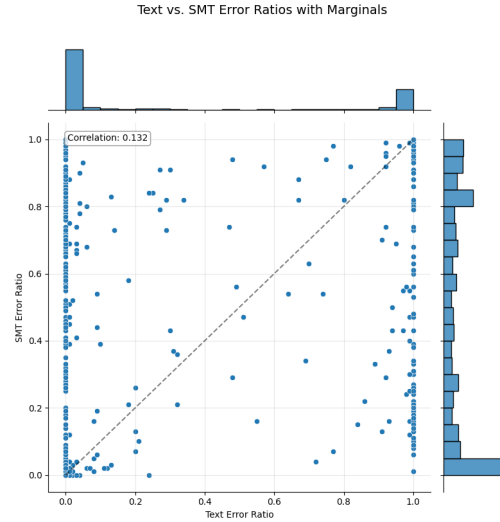


Figure 15: SMT vs. Text Error Ratio Analysis for Gemini Flash 2.0: Depicts less-calibrated SMT generation, evidenced by a weaker correlation between SMT and Text error patterns.

972 The figures juxtapose SMT versus Text error ratios (with marginals) for o3-mini (Fig. 14) and
 973 Gemini Flash 2.0 (Fig. 15); the Text error ratio is defined as the proportion of incorrect direct
 974 textual answers from the LLM per question out of the many samples, while the SMT error ratio is
 975 the proportion of incorrect answers derived from its SMT-LIB formalizations. O3-mini exhibits a

notable correlation between its SMT and Text error distributions, characteristic of a *well-calibrated SMT generation process* where formalization errors tend to align with textual reasoning errors. In contrast, Gemini Flash 2.0 shows a weaker correlation, suggesting its SMT generations may introduce errors or exhibit patterns less consistently coupled with its textual output, indicative of *poorer calibration*. This comparative error ratio analysis is valuable for assessing the fidelity of an LLM’s autoformalization. Strong SMT-Text error correlation implies that the SMT modality can be a more reliable indicator of the LLM’s general reasoning tendencies for a problem, making SMT-derived uncertainty metrics potentially more transferable. Poor correlation, however, signals a divergence between textual reasoning and formalization, cautioning against using SMT outputs as direct proxies without careful consideration of modality-specific error sources and motivating efforts towards better SMT-Text reasoning alignment.

F Qualitative Analysis

Beyond quantitative uncertainty metrics, the PCFG framework, by its nature of parsing and structuring program ensembles, lends itself to a nuanced qualitative analysis of LLM-generated formal artifacts. Initial explorations can focus on broad characteristics such as the distribution of SMT-LIB sorts (datatypes) employed or the prevalent logical fragments (e.g., ‘QF_LIA’, ‘QF_AUFBV’) selected by the LLM for a given problem class. However, a more profound understanding of an LLM’s formalization strategy emerges from a detailed examination of substructures, like the `assert` statements, which constitute the semantic core of an SMT program by stipulating the conditions and axioms for the solver. Our PCFG-based analysis of these assertions, and the logical architectures therein, reveals critical patterns in how LLMs attempt to translate natural language problem specifications into rigorous, machine-interpretable logic.

When an LLM generates multiple SMT program samples for a single natural language input, the per-problem induced PCFG captures a distribution over grammatical structures. This distribution inherently models the LLM’s normative formalization pathways alongside its idiosyncratic variations, particularly in the construction of `assert` statements and their nested logical terms—including quantifiers (`forall`, `exists`), logical connectives (\Rightarrow , `and`, `or`, `not`), and predicate applications. Analyzing the probabilities and diversity of production rules within these PCFGs allows for the identification and interpretation of several key types of divergences and tendencies in formal specification:

Formalization Aliasing and Representational Stability A core aspect of a problem specification may elicit syntactically diverse, yet ideally logically equivalent, `assert` statements across an ensemble of LLM generations. For instance, an implication $A \Rightarrow B$ might be directly asserted or rendered as $\neg A \vee B$. The PCFG reflects such syntactic polymorphism through multiple, lower-probability rule sequences mapping to the same underlying semantic constraint. A high degree of such variability for asserting fundamental problem axioms often signals the LLM’s lack of a converged or canonical formalization strategy, potentially indicating uncertainty or representational underspecification for that particular logical construct.

Variance in Logical Decomposition and Structural Complexity The PCFG rule sets unveil the LLM’s implicit preferences regarding the structural complexity and granularity of asserted logical terms. For a given problem, some SMT samples might employ deeply nested quantifiers and connectives within a monolithic `assert` statement. In contrast, other samples might exhibit a preference for flatter, more direct assertions or decompose a complex axiom into several simpler, conjoined `assert` statements. This divergence in logical decomposition strategies is captured by differing rule probabilities and derivation depths within the PCFG, pointing to variations in the LLM’s approach to abstraction and information chunking during the formalization process.

Identification of Atypical or Anomalous Assertions Occasionally, an LLM may generate `assert` statements possessing highly unusual or infrequent syntactic structures relative to the typical formalizations observed for a given problem context or across a dataset. The PCFG methodology inherently highlights these as low-probability production rules or derivations. Qualitative inspection of the SMT code corresponding to these rare assertion patterns can uncover unique, potentially innovative, or conversely, flawed and overly convoluted ways the LLM attempts to axiomatize specific constraints, offering insights into its error modes or its capacity for novel formal expression.

1028 **Semantic Divergence in Axiomatization** More critically, divergences can be semantic rather than
 1029 merely syntactic, leading to logically distinct problem formalizations from the same natural language
 1030 input. Such semantic drift often manifests as significantly different asserted terms within `assert`
 1031 statements, pointing to LLM misinterpretation, unresolved ambiguity, or flaws in its inferential
 1032 reasoning. For example, if an input "All engineers use LaTeX" is ambiguously formalized, one
 1033 SMT sample might correctly assert `(forall ((x Engineer)) (usesLaTeX x))`, while a seman-
 1034 tically divergent sample might erroneously assert `(forall ((x User)) (implies (usesLaTeX
 1035 x) (isEngineer x)))`. The PCFG rules governing the predicates, variables, and logical structure
 1036 of terms within these assertions would markedly differ, directly reflecting this semantic incongruity
 1037 and providing a diagnostic trace.

1038 **Fidelity in Representing Ground Facts** For declarative factual statements present in the input (e.g.,
 1039 "Constantine is a logician"), the LLM is expected to consistently assert the corresponding ground fact
 1040 in a stable manner. If the PCFG reveals multiple, conflicting, or unstable rule applications for asserting
 1041 properties of specific entities (e.g., some derivations asserting `(isLogician constantine)` while
 1042 others, for the same conceptual input fact, generate `(isPhilosopher constantine)` or vary the
 1043 predicate structure), this indicates a deficiency in the LLM’s fidelity in extracting and consistently
 1044 formalizing elementary factual information, pointing to potential grounding issues.

Program 1	Program 2	Program 3	Program 4
<pre> (set-logic QF_LIA) (declare-const relation_count_G Int) (declare-const relation_count_J Int) (assert (> relation_count_G relation_count_J)) (assert (>= relation_count_G 1)) (assert (>= relation_count_J 0)) (check-sat) (get-model) </pre>	<pre> (set-logic QF_LIA) (declare-const GC Int) (declare-const JC Int) (assert (> GC 0)) (assert (> JC 0)) (assert (> GC JC)) (assert (>= GC 10)) (assert (>= JC 1)) (check-sat) (get-model) </pre>	<pre> (set-logic QF_LIA) (declare-fun people_genghis () Int) (declare-fun people_caesar () Int) (assert (>= people_genghis 0)) (assert (>= people_caesar 0)) (assert (> people_genghis people_caesar)) (assert (>= people_genghis 1000000)) (assert (<= people_caesar 500000)) (check-sat) (get-model) </pre>	<pre> (set-logic QF_LIA) (declare-const KN Int) (declare-const CA Int) (assert (> KN CA)) (assert (= KN 16)) (assert (= CA 1)) (check-sat) (get-model) </pre>

Table 18: Divergent LLM Formalizations of a StrategyQA Problem: Sample SMT-LIB outputs illustrating varied axiomatization strategies, with `assert` statements highlighted (blue). Such variations are central to the qualitative PCFG analysis discussed.