MEANING BEFORE WORDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting words and preserving meaning are different tasks. We study this gap in a finite language model where many strings express one meaning. The probability chain rule can already represent any latent semantic generator; the obstruction lies in the decisions and resource allocation built around it. For every fixed temperature below one, we construct a full-support binary model whose semantic error tends to zero, while both token-level and sequence-level temperature sampling have error tending to one. Even exact sequence maximization fails. We give a sharp finite-length bound for the local effect. Under a shared memory limit, an optimal next-token predictor can discard an available semantic bit; we compute the exact cost of protecting it. A sharp log-loss bound separates semantic guarantees from length-normalization artifacts. Finally, a meaning-first interface makes decisions independent of verbal form when its decoder is faithful. The results separate semantic decision-making from its autoregressive language realization.

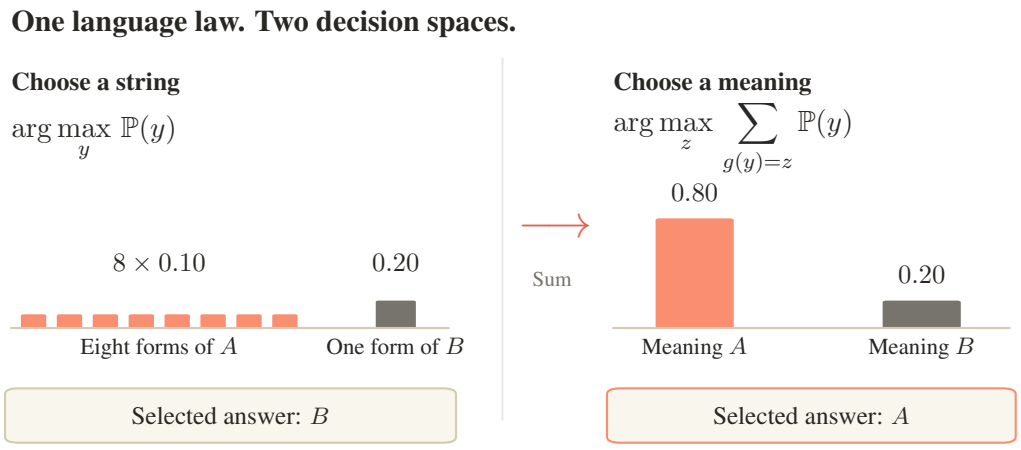


Figure 1: The same law can favor different answers. Eight expressions of A carry mass 0.10 each; one expression of B carries 0.20. String MAP selects B, while semantic MAP selects A. This illustrative law isolates form competition; it does not assume that model probabilities certify truth.

1 INTRODUCTION

A language model must decide both *what to say* and *how to say it*. These are not the same decision. "Yes," "That is correct," and a longer explanation may express the same answer. A probability model assigns mass to each string. A user often cares about the answer they share. Maximizing a string probability can therefore select a different answer from maximizing the total probability of its meaning. This is the surface form competition identified by Holtzman et al. (2021).

A natural response is to choose a meaning first and then express it. This suggests a latent variable z: first form an idea, then generate words conditional on that idea. The suggestion is useful, but a common argument for it is not. Integrating out z does not itself destroy the joint distribution. An exact next-token model can reproduce every observable law produced by a latent semantic model. Nor does a large set of possible wordings, by itself, force a finite model to forget meaning.

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

Table 1: The assumptions determine the conclusion. Each row asks a different question; no row is an unconditional impossibility theorem for autoregression.

Question	Required assumption	Conclusion
Does integrating out z lose meaning?	Exact posterior-weighted conditionals	The joint law is unchanged.
Can sharper decoding hurt?	Fixed cooling; our full-support binary family	Semantic error can rise from near zero to near one.
Does small token loss suffice?	Vanishing mean excess loss; growing length	Semantic error can still approach one.
Can form displace meaning?	Theorem 3's interface; $1 \le C \le m$, $0 < \lambda < 1$	Every exact predictive optimum drops the semantic bit.
What does separation guarantee?	An explicit semantic decision; a faithful decoder	Form changes leave the semantic law unchanged.

The right question is narrower: *which operations on a token model preserve the semantic task?* We separate three operations. Decoding changes the generated distribution. Averaging log loss changes the scale of its error certificate. Compressing the context changes the information available for later prediction. None is part of the probability chain rule, so each needs its own analysis.

Our strongest separation needs no training error. For a model over two tokens and two meanings, ordinary samples are almost always correct. Yet any fixed cooling makes the wrong meaning almost certain as length grows. The wrong string is also the unique global mode, so exact sequence search cannot fix the example. A sharp path bound identifies the critical output-length scale of this family.

Our second result gives a precise sense in which form can crowd out meaning. A bounded representation must serve both form queries and a semantic query. When each form query is weighted more heavily, every optimum of the next-token objective forgets the semantic bit. Reserving one bit for meaning repairs this failure, but has a nonzero, exactly determined cost in predictive loss. Thus, semantic protection is a choice of objective or constraint, not free extra capacity.

2 A SMALL MODEL OF MEANING AND FORM

Start with the token law. Fix a prompt x, a finite vocabulary $\mathcal{V}$, and a length T. A language model defines a joint law on $Y = (Y_1, \dots, Y_T) \in \mathcal{Y} = \mathcal{V}^T$. The probability chain rule gives

$$\mathbb{P}(y \mid x) = \prod_{t=1}^{T} \mathbb{P}(y_t \mid x, y_{<t}). \tag{1}$$

This probability identity places no restriction on how conditionals are computed (Cover & Thomas, 2006, Chapter 2). End-of-sequence padding handles bounded variable-length output. We suppress the fixed prompt when unnecessary. All logarithms are base two, except the natural logarithm $\ln$.

Specify what the task counts as meaning. A known map $g_x : \mathcal{Y} \to \mathcal{Z}$ assigns a semantic label to each complete string. It may record an answer, a logical claim, or a structured plan. Write $Z = g_x(Y)$ and $\mathcal{Y}_z = \{y : g_x(y) = z\}$. We use *form* for distinctions within $\mathcal{Y}_z$. This includes wording and syntactic realization, but is not a claim that all linguistic syntax is independent of meaning. The partition is part of the task specification, not something our theorems learn.

Following semantic probability aggregation (Kuhn et al., 2023), define

$$\pi(z \mid x) = \sum_{y \in \mathcal{Y}_z} \mathbb{P}(y \mid x), \quad r(y \mid x, z) = \mathbb{P}(y \mid x, Z = z). \tag{2}$$

For $\pi(z \mid x) > 0$, π selects *meaning* and r selects *expression*; counts need not predict mass.

Proposition 1 (Marginalization does not lose the language law)**.** *The faithful generator $Z \sim \pi(\cdot \mid x)$, $Y \sim r(\cdot \mid x, Z)$ induces the same observable law as the following mixture:*

$$\mathbb{P}(y \mid x) = \sum_z \pi(z \mid x) r(y \mid x, z). \tag{3}$$

At every positive-probability prefix u, the posterior-weighted next-token law is

$$\mathbb{P}(a \mid x, u) = \sum_z \mathbb{P}(z \mid x, u) r(a \mid x, u, z). \tag{4}$$

Conversely, any law $\mathbb{P}$ and any semantic map g_x admit this faithful latent representation.

The weights in (4) are the *updated posterior*, not the original prior $\pi(z \mid x)$. Reusing the prior at every token generally creates a different process. Exact autoregression can carry semantic uncertainty through its hidden computation; the proposition does not require an explicit, externally readable z. It also does not establish that a learned representation is identifiable or easy to compute.

Form has a size, but size is not yet pressure. The entropy chain rule gives

$$H(Y \mid x) = H(Z \mid x) + H(Y \mid Z, x). \tag{5}$$

Here entropy measures uncertainty in bits (Cover & Thomas, 2006, Chapter 2). The last term measures uncertainty about expression after meaning is fixed. It can be large while $H(Z \mid x)$ is small. This identity alone implies no loss of meaning: an unrestricted model can represent both terms exactly. Pressure appears only after a decision rule, a loss scale, or a resource constraint is specified.

3 When confidence in words defeats confidence in meaning

3.1 Aggregation and sharpening are different operations

A useful semantic decision should not change merely because one expression is split into several equivalent forms. For example, replacing an expression of mass p by two synonymous expressions of mass $p/2$ leaves its total meaning probability unchanged. We call this a *form refinement*. The comparison concerns laws on semantic labels, not equality of the resulting strings.

To sharpen a distribution, apply a fixed scalar function separately to each probability and renormalize. Let $w : [0, 1] \to [0, \infty)$ be continuous, with $w(0) = 0$ and $w(p) > 0$ for $p > 0$. Set

$$(T_w \mathbb{P})(y) = \frac{w(\mathbb{P}(y))}{\sum_v w(\mathbb{P}(v))}. \tag{6}$$

Here and below a sum over v in a complete-string law is over strings, not tokens.

Proposition 2 (The only universally refinement-compatible pointwise rule)**.** *Transforming and aggregating commute for every finite law $\mathbb{P}$ and every semantic partition if and only if $w(p) = cp$ for some $c > 0$. Thus ordinary sampling is the only universally compatible rule within this class.*

This characterizes additivity; it does not require every decoder to sample. A decoder that knows g_x can aggregate before making a task-specific decision. The obstruction concerns *blind* nonlinear sharpening, which cannot respect every semantic partition. Appendix B gives the proof.

For inverse temperature $\beta = 1/\tau$, sequence-level sharpening uses $w(p) = p^\beta$, giving

$$\mathbb{Q}_\beta^{\text{seq}}(Z = z) = \frac{\pi(z)^\beta A_\beta(z)}{\sum_j \pi(j)^\beta A_\beta(j)}, \quad A_\beta(z) = \sum_{y \in \mathcal{Y}_z} r(y \mid z)^\beta. \tag{7}$$

This derived expression is proved in Appendix B. The factor $A_\beta(z)$ depends on form. If z has n_z equally likely expressions, it equals $n_z^{1-\beta}$. For $\beta > 1$, a meaning spread across more such expressions is penalized. Unequal expression probabilities require A_β, not a count alone.

Token temperature transforms each prefix law (Stewart, 2016; Kempton & Burrell, 2025):

$$\mathbb{Q}_\beta^{\text{loc}}(a \mid u) = \frac{\mathbb{P}(a \mid u)^\beta}{\sum_{b \in \mathcal{V}} \mathbb{P}(b \mid u)^\beta}. \tag{8}$$

Local and sequence-level sharpening generally differ: the former multiplies prefix normalizers into a path-dependent correction. We keep them distinct and show that *both* can fail on one task.

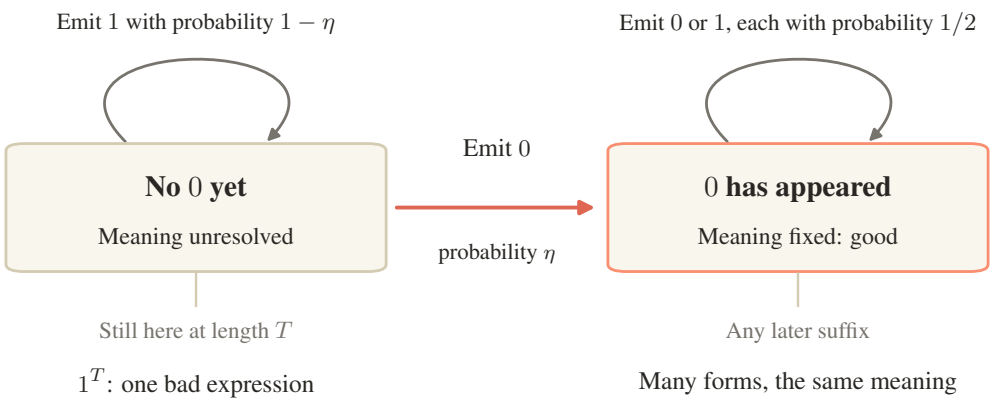


Figure 2: An exact binary model with many good expressions and one bad expression. Local cooling suppresses each opportunity to reach the good class; global cooling penalizes mass spread over its forms. The construction isolates a semantic decision from its subsequent realization.

3.2 A FULL-SUPPORT BINARY SEPARATION

In our binary language, any string containing 0 expresses good; only 1^T expresses bad. Until the first 0, the model emits 0 with probability η and 1 with probability $p = 1 - \eta$. Later fair bits vary only form. The law has full support; one binary state bit tracks whether a zero has already occurred.

Theorem 1 (Accurate samples, incorrect confident decoding)**.** *Fix any $\beta > 1$ and choose $a \in (1/\beta, 1)$. For sufficiently large T, set $\eta = T^{-a}$ in the binary model above. As $T \to \infty$,*

$$\mathbb{P}(\mathsf{bad}) \longrightarrow 0, \quad \mathbb{Q}_\beta^{\mathrm{loc}}(\mathsf{bad}) \longrightarrow 1, \quad \mathbb{Q}_\beta^{\mathrm{seq}}(\mathsf{bad}) \longrightarrow 1. \tag{9}$$

For all sufficiently large T, both greedy decoding and exact sequence MAP output bad*. In contrast, sharpening the two semantic masses rather than individual string probabilities gives*

$$\mathbb{Q}_\beta^{\mathrm{sem}}(\mathsf{bad}) := \frac{\mathbb{P}(\mathsf{bad})^\beta}{\mathbb{P}(\mathsf{bad})^\beta + \mathbb{P}(\mathsf{good})^\beta} \longrightarrow 0. \tag{10}$$

What the theorem says. An exact model can be reliable under sampling and unreliable under confidence-seeking decoding. Neither approximate inference nor a growing vocabulary causes this failure. Good continuations together carry almost all the mass, yet none individually beats the bad string. This gives a full-support binary example with almost complete error reversal.

Why the local failure grows. The exact bad mass is $(1-\eta)^T$. Ordinary sampling has T chances to take a good transition of probability η. Cooling reduces that transition to order η^β. The choice of a makes $T\eta$ diverge but $T\eta^\beta$ vanish. These competing scales, established precisely in Appendix C, explain the reversal directly, without invoking a vague or unmeasured loss of semantic information.

At $T = 1000$, $\eta = 0.01$, and $\tau = 0.5$, the reference bad probability is below 0.00005, but local cooling raises it above 0.90 and global cooling above 0.999. These values come from exact formulas, not experiments on trained models. The reference is accurate, not error-free: positive-temperature sharpening preserves support and cannot create mass on an impossible event.

3.3 A SHARP FINITE-LENGTH ENVELOPE

We also give a finite-length bound: maximize local amplification along a path whose chosen token is never less likely than its binary alternative. No local preference is required of other paths.

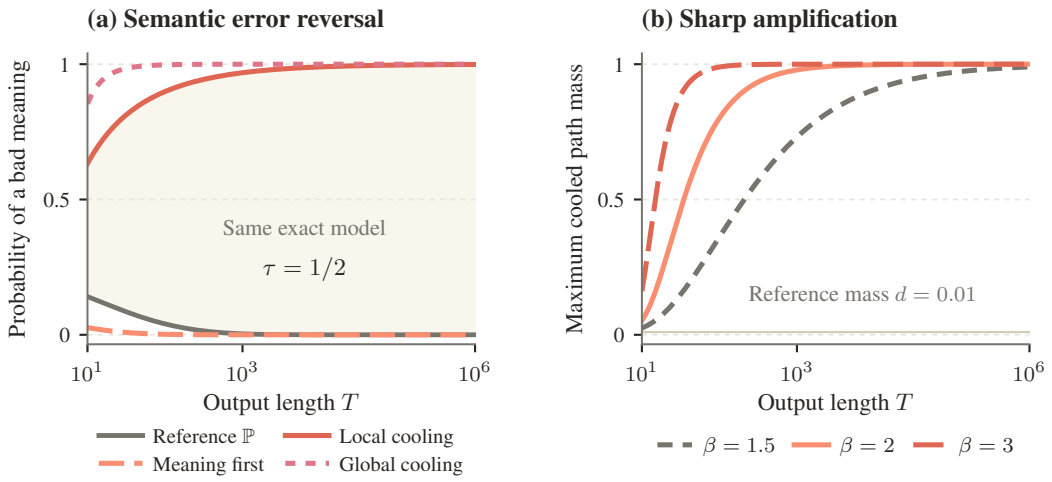


Figure 3: Analytic consequences, not empirical evaluations. Left: exact semantic-error formulas for Theorem 1, with $\beta = 2$ and $\eta = T^{-3/4}$. Right: the sharp path envelope for fixed reference bad mass $d = 0.01$; all plotted horizons satisfy $d \geq 2^{-T}$. Increasing length permits stronger amplification even when the original path mass is unchanged.

Theorem 2 (Sharp amplification on a locally preferred path)**.** *For a binary law, let* $p_t = \mathbb{P}(1 \mid 1^{t-1}) \in [1/2, 1]$ *and* $d = \mathbb{P}(1^T) = \prod_t p_t$*. For* $\beta > 1$*, the cooled path mass satisfies*

$$\mathbb{Q}_\beta^{\mathrm{loc}}(1^T) \leq B_{\beta,T}(d) := \left[1 + \left(d^{-1/T} - 1\right)^\beta\right]^{-T}, \quad d \in [2^{-T}, 1]. \tag{11}$$

Equality holds exactly when $p_1 = \cdots = p_T = d^{1/T}$*, and is attained with full support for each* $d < 1$ *in this interval. The tied fair path is allowed, but full support excludes* $d = 1$*.*

At fixed path mass, equal confidence across steps gives the strongest amplification. The bound identifies the critical dependence on output length. If $L_T = -\ln d_T > 0$ and $L_T = o(T)$, then

$$-\ln B_{\beta,T}(d_T) \sim \frac{L_T^\beta}{T^{\beta-1}}. \tag{12}$$

Thus d_T may vanish while its sharp cooled upper bound tends to one. Appendix D proves the bound, endpoints, and critical scale. This worst-case result need not describe typical applications.

4 What Token Log Loss Does and Does Not Certify

The previous section kept the model exact and changed decoding. We now ask what log loss certifies about an approximate model $\mathbb{Q}$ relative to a reference law $\mathbb{P}$. Define mean excess token loss by

$$\epsilon_T(\mathbb{P}, \mathbb{Q}) := \frac{1}{T}\mathbb{E}_{Y \sim \mathbb{P}(\cdot)}\big[-\log \mathbb{Q}(Y) + \log \mathbb{P}(Y)\big] = \frac{1}{T} D_{\mathrm{KL}}(\mathbb{P} \| \mathbb{Q}). \tag{13}$$

This standard log-loss identity follows from the chain rule (Cover & Thomas, 2006, Chapter 2). It measures excess loss relative to the true next-token predictor, rather than raw cross-entropy.

Likelihood controls meaning at the correct scale. The same fixed map gives the KL decomposition

$$D_{\mathrm{KL}}(\mathbb{P} \| \mathbb{Q}) = D_{\mathrm{KL}}(\mathbb{P}_Z \| \mathbb{Q}_Z) + \mathbb{E}_{Z \sim \mathbb{P}_Z(\cdot)} D_{\mathrm{KL}}(\mathbb{P}(\cdot \mid Z) \| \mathbb{Q}(\cdot \mid Z)). \tag{14}$$

Both terms are nonnegative. Small *sequence* KL controls semantic KL: unconstrained likelihood fitting need not trade them off. Appendix E derives this identity, including zero-probability cases.

Proposition 3 (A sharp semantic-error certificate). *Let E be the event that the output has an incorrect semantic label. If $\mathbb{P}(E) = 0$ and $\epsilon_T(\mathbb{P}, \mathbb{Q}) \le \epsilon$, then ordinary sampling from $\mathbb{Q}$ satisfies*

$$\mathbb{Q}(E) \le 1 - 2^{-T\epsilon}. \tag{15}$$

For every finite T and $\epsilon \ge 0$, the bound is attained whenever a correct and an incorrect output are available. There also exist binary examples with $\epsilon_T \to 0$ but $\mathbb{Q}(E) \to 1$ as $T \to \infty$.

The last claim warns against normalization, not likelihood itself. Take $T-1$ fair form bits and one final semantic bit. The reference ends in 0; the learned law ends in 1 with probability $1-2^{-\sqrt{T}}$. Only that final prediction incurs excess loss: its sequence KL is $\sqrt{T}$, but the average is $1/\sqrt{T}$. Correctly modeled form dilutes the reported average loss without fixing the decisive semantic error.

The certificate needs truthful reference support and ordinary sampling. Real corpora may not be truthful; cooling changes the law. Other hallucination bounds make different assumptions (Kalai & Vempala, 2024; Kalai et al., 2025). Low mean loss and latent variables alone cannot certify facts.

5 When predictive memory chooses form over meaning

To turn "crowding out" into a theorem, we need a shared budget and a reason to favor form. Consider a context containing independent fair bits $U_1, \dots, U_m, Z$. The U_i specify form choices; Z specifies a semantic fact. An encoder sees them all and retains a state H with at most 2^C values. The original context is then unavailable. Only afterward is a query J, independent of all source bits, revealed: $J = i$ asks for U_i, while $J = 0$ asks for Z. Every answer is a single binary token.

Both parties may share a public seed R, independent of the source and query jointly. The encoder may use independent private randomness but cannot see J. The decoder receives only (H, R, J), excluding uncharged buffers, correlated seeds, and teacher-forced answers. The query law is

$$\mathbb{P}(J = i) = \frac{1}{m+\lambda}\ (1 \le i \le m), \quad \mathbb{P}(J = 0) = \frac{\lambda}{m+\lambda}, \quad 0 < \lambda < 1. \tag{16}$$

Each form query outweighs the semantic query by assumption, not merely because more forms exist.

For a fixed encoder, the posterior decoder minimizes expected next-token log loss, giving

$$\mathcal{L}(H) = \frac{\sum_{i=1}^{m} H(U_i \mid H, R) + \lambda H(Z \mid H, R)}{m+\lambda}. \tag{17}$$

This specializes Bayes log loss (Cover & Thomas, 2006); see Appendix F. Mutual information $I(Z; H \mid R)$ measures the semantic bit's uncertainty resolved by the retained state.

Theorem 3 (The exact cost of semantic retention). *Let $1 \le C \le m$ be an integer. The optimum over all admissible encoders and the semantic-retention cost for any such encoder obey*

$$\mathcal{L}_* = \frac{m+\lambda-C}{m+\lambda}, \quad \mathcal{L}(H) - \mathcal{L}_* \ge \frac{1-\lambda}{m+\lambda} I(Z; H \mid R). \tag{18}$$

With source- and query-independent public time sharing, the bound is attained at every retention level in $[0, 1]$. Every exact optimum has $I(Z; H \mid R) = 0$ and semantic Bayes error $1/2$. Protecting the entire semantic bit costs exactly $(1-\lambda)/(m+\lambda)$ in loss, at unchanged capacity.

The optimum spends all C bits on form. Protecting Z instead leaves $C-1$ bits for form. The theorem does not assume this coordinate-wise coding structure: its converse covers arbitrary joint and randomized encoders. The two simple codes only establish attainability. This is an explicit task-dependent compression tradeoff, in the sense of the information bottleneck (Tishby et al., 2000).

Semantic protection can be cheap and still lose under the objective. For $m = 99$, $C = 8$, and $\lambda = 1/2$, reserving one bit for Z reduces semantic error from $1/2$ to zero, costing only $1/199$ bits per query. Yet every exact optimum leaves Z unprotected. The example separates a penalty's *size* from the lost information's *importance*; it says nothing about neural-network loss landscapes.

With decoder loss at most $\mathcal{L}_* + \Delta$, the best semantic error $e_Z \in [0, 1/2]$ must satisfy

$$h_2(e_Z) \ge \max\left\{0, 1 - \frac{(m+\lambda)\Delta}{1-\lambda}\right\}, \tag{19}$$

where $h_2(e) = -e \log e - (1-e)\log(1-e)$ is binary entropy. This follows from (18) and binary Fano's inequality (Cover & Thomas, 2006, Section 2.10); the full argument is in Appendix F. We claim sharpness for the information–loss tradeoff, not for this error-rate relaxation.

Can more generated tokens repair the loss? A long verbal explanation may perform useful computation, but it cannot reveal a source bit that its starting information has discarded. This distinction follows from data processing (Cover & Thomas, 2006, Section 2.8). In the delayed-query model, it yields an output-level consequence rather than only a statement about a hidden state.

Corollary 1 (Generated text cannot restore erased information)**.** *From (H, R) and fresh randomness jointly independent of (U, Z, H, R), generate an almost surely finite transcript W with no length or computation bound. It adds no semantic information beyond the retained state:*

$$H(Z \mid H, R, W) = H(Z \mid H, R). \tag{20}$$

The best semantic error using (H, R, W) is exactly e_Z, the error using (H, R) alone. In particular, every optimal encoder in Theorem 3 remains at semantic error $1/2$ after any such transcript.

Appendix F uses conditional independence. Any autoregressive generator is allowed; the decoder may read the full transcript. No new source-dependent evidence is permitted: neither rereading the erased context nor consulting an external source. Retaining that bit prevents this failure.

This is an information limit, not an argument against chain-of-thought computation. Additional tokens can make already retained information easier for a computationally bounded decoder to use. The corollary compares optimal decoders without computational limits. The point is to separate *recovering what was never lost* from *reconstructing what is no longer available*.

The restrictions matter. If $\lambda = 1$, forgetting is no longer forced; if $C \geq m + 1$, all bits fit. If the original context remains accessible, this particular bottleneck is absent. The theorem is therefore not a lower bound on every Transformer hidden vector or KV cache. It shows that next-token optimality, by itself, does not select semantic retention under a shared, specified resource limit.

6 A SEMANTIC INTERFACE, NOT A NEW CHAIN RULE

Choose the decision at the level of its loss. Additionally suppose the correct label Z^* has posterior $\mathbb{P}(Z^* = z \mid x) = \pi(z \mid x)$; this does not follow from $\pi = \mathbb{P}_Z$ alone. A fixed string y is then correct with probability $\pi(g_x(y))$. Semantic zero–one loss selects the class with largest *total* mass, then any expression within it. String MAP instead selects the largest individual probability. Appendix B gives this standard Bayes rule (Cover & Thomas, 2006; Eikema & Aziz, 2020).

The semantic decision also tolerates bounded errors in aggregation. Let $\tilde{\pi}$ be an approximate semantic law with total variation distance at most ε from π, where total variation is half the sum of absolute probability differences. If $\tilde{z} \in \arg\max_z \tilde{\pi}(z)$ and $z_\star \in \arg\max_z \pi(z)$, then

$$\pi(z_\star) - \pi(\tilde{z}) \leq 2\varepsilon. \tag{21}$$

Appendix G proves this bound. Only error in semantic masses enters, not the number of synonymous strings. The result neither supplies those masses nor ensures that the posterior is factually informed.

Protect the decision, then vary its expression. A planner draws $\hat{Z} \sim \hat{\pi}(\cdot \mid x)$; a decoder draws $Y \sim \mathbb{K}(\cdot \mid x, \hat{Z})$ using fresh randomness. A *faithful* interface has $g_x(Y) = \hat{Z}$ almost surely, so wording changes preserve the planned label. A named hidden vector does not imply fidelity.

Proposition 4 (Separation of decision and realization)**.** *A faithful decoder induces semantic law $\hat{\pi}$, independently of its within-class laws $\mathbb{K}(\cdot \mid x, z)$. More generally, let $\delta, \rho \in [0, 1]$, let $z^*(x)$ be the target, and let planner error be at most δ. Suppose every reachable (x, z) obeys*

$$\mathbb{K}\big(g_x(Y) \neq z \mid x, z\big) \leq \rho. \tag{22}$$

The planner and decoder errors then give the following sharp bound on total semantic error:

$$\mathbb{P}\big(g_X(Y) \neq z^*(X)\big) \leq \delta + (1-\delta)\rho. \tag{23}$$

Adding faithful forms introduces no length-dependent term; the bound remains sharp.

The uniform condition in (22) is the real requirement. A hidden vector called "thought" need not satisfy it, and a decoder may still mistranslate a correct plan. Moreover, ρ itself may grow with length in an actual implementation. The proposition isolates the two failures so that they can be measured or bounded separately; it does not make either failure disappear by definition.

The decoder can remain autoregressive. For a planned label z with $\pi(z) > 0$ and a prefix u with $\mathbb{P}(u) > 0$, set $h_z(u) = \mathbb{P}(g_x(Y) = z \mid u)$. An exactly faithful autoregressive decoder is

$$\mathbb{K}_z(a \mid u) = \mathbb{P}(a \mid u)\frac{h_z(ua)}{h_z(u)} \quad \text{when } h_z(u) > 0. \tag{24}$$

Appendix G proves normalization and faithfulness, so the construction needs no departure from autoregression. It is not a cheap algorithm: h_z may require reasoning over completions. Efficient local scores do not supply semantic continuation masses for free (Lin et al., 2021).

What can be claimed about reasoning cost? A counting bound isolates serialization cost. A checkpoint must retain one of N semantic states and one of M prescribed form states; all NM pairs are admissible. It passes at most an integer $C \geq 0$ hidden bits and exactly L symbols from an alphabet of size $A \geq 2$, with no other source-dependent state. Lossless recovery requires

$$2^C A^L \geq NM; \quad \text{semantic recovery alone requires } 2^C A^L \geq N. \tag{25}$$

Arbitrary joint coding attains both bounds (Appendix H). Preserving a prescribed verbal form may therefore require additional communication that the next semantic update does not need.

This is not a universal speedup theorem. A decoder may choose fresh wording rather than preserve old wording, the full prefix may already be available, and latent computation also costs time and memory. Final text still has to be emitted. The defensible design principle is to avoid *requiring* every internal semantic update to cross an unnecessary language interface. Pause tokens, continuous latent reasoning, and concept-level models explore related choices (Goyal et al., 2024; Hao et al., 2025; LCM team et al., 2024); our results do not certify their learned representations.

7 Relation to prior work and scope

Meaning-aware decoding is not a new idea. Stewart (2016) identified semantic distortion from temperature and proposed semantic temperature sampling. Surface form competition (Holtzman et al., 2021; Wiegreffe et al., 2023), semantic entropy (Kuhn et al., 2023), and critiques of MAP decoding (Eikema & Aziz, 2020) also separate meaning from strings. Proposition 2 gives a structural explanation. We add the full-support binary reversal and Theorem 2's sharp path envelope.

Local normalization is only part of the problem. Kempton & Burrell (2025) analyze local distortion and quality–diversity objectives. We distinguish local from sequence-level temperature, but evaluate a supplied semantic event rather than likelihood as a quality proxy. In Theorem 1, global normalization and exact sequence MAP still fail. Normalization alone cannot guarantee meaning.

Learning and representation require their own assumptions. Bachmann & Nagarajan (2024) study teacher-forcing shortcuts in planning, while Rohatgi et al. (2025) establish barriers under misspecification and a specified next-token learning interface. Our decoding construction is already exact; our memory theorem uses a different, fully charged delayed-query interface. Positive identifiability results also show that next-token training can encode latent concepts under additional assumptions (Liu et al., 2026). We do not claim that semantic representations cannot emerge.

The semantic map is an assumption, not a discovery. Corielli (2026) distinguish latent circumstances, observed text, and learned text laws. We study a known finite semantic partition under explicit interventions. We do not solve partition learning, world grounding, or open-ended fact checking. A wrong partition or uninformed planner may produce fluent, consistently wrong text. The theory makes no universal claim about the temporal order of human thought and speech.

8 CONCLUSION

Next-token prediction can represent a semantic generator exactly, yet representation alone does not guarantee meaning. Sharpening may prefer a wrong answer; averaged loss may hide decisive errors; a shared budget may favor semantic forgetting. Each failure requires assumptions beyond the chain rule. A meaning-first interface makes the semantic decision explicit and requires faithful decoding. Token prediction can still serve that interface, but need not be the whole solution.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2296–2318. PMLR, 2024. URL https://proceedings.mlr.press/v235/bachmann24a.html.

Francesco Corielli. When is next-token prediction useful? marginalization, ergodicity, mixture identifiability, local sufficiency, RAG, tools, and programming. arXiv:2605.23278v2, 2026. URL https://arxiv.org/abs/2605.23278v2.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. doi: 10.1002/047174882X. URL https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of COLING*, pp. 4506–4520, 2020. doi: 10.18653/v1/2020.coling-main.398. URL https://aclanthology.org/2020.coling-main.398/.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02226v3.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2412.06769v4. Revised version: arXiv:2412.06769v4, 2026.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of EMNLP*, pp. 7038–7051. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.564. URL https://aclanthology.org/2021.emnlp-main.564/.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, 2024. URL https://arxiv.org/abs/2311.14648v3.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. arXiv:2509.04664v1, 2025. URL https://arxiv.org/abs/2509.04664v1.

Tom Kempton and Stuart Burrell. Local normalization distortion and the thermodynamic formalism of decoding strategies for large language models. In *Findings of EMNLP*, pp. 22216–22231. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1210. URL https://aclanthology.org/2025.findings-emnlp.1210/.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09664v3.

LCM team, Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R. Costa-jussà, David Dale, Hady Elsahar, Kevin Heffernan, João Maria Janeiro, Tuan Tran, Christophe Ropers, Eduardo Sánchez, Robin San Roman, Alexandre Mourachko, Safiyyah Saleem, and Holger Schwenk. Large concept models: Language modeling in a sentence representation space. arXiv:2412.08821v2, 2024. URL https://arxiv.org/abs/2412.08821v2.

Chu-Cheng Lin, Aaron Jaech, Xin Li, Matthew R. Gormley, and Jason Eisner. Limitations of autoregressive models and their alternatives. In *Proceedings of NAACL-HLT*, pp. 5147–5173. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.405. URL https://aclanthology.org/2021.naacl-main.405/.

Yuhang Liu, Dong Gong, Yichao Cai, Erdun Gao, Zhen Zhang, Biwei Huang, Mingming Gong, Anton van den Hengel, and Javen Qinfeng Shi. I predict therefore i am: Is next token prediction enough to learn human-interpretable concepts from data? arXiv:2503.08980v4, 2026. URL https://arxiv.org/abs/2503.08980v4.

Dhruv Rohatgi, Adam Block, Audrey Huang, Akshay Krishnamurthy, and Dylan J. Foster. Computational-statistical tradeoffs at the next-token prediction barrier: Autoregressive and imitation learning under misspecification. arXiv:2502.12465v1, 2025. URL https://arxiv.org/abs/2502.12465v1.

Russell Stewart. Maximum likelihood decoding with RNNs—the good, the bad, and the ugly. Stanford Natural Language Processing Group blog, April 2016. URL https://www-nlp.stanford.edu/blog/maximum-likelihood-decoding-with-rnns-the-good-the-bad-and-the-ugly/.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. arXiv:physics/0004057, 2000. URL https://arxiv.org/abs/physics/0004057. Manuscript dated September 1999.

Sarah Wiegreffe, Matthew Finlayson, Oyvind Tafjord, Peter Clark, and Ashish Sabharwal. Increasing probability mass on answer choices does not always improve accuracy. In *Proceedings of EMNLP*, pp. 8392–8417. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.522. URL https://aclanthology.org/2023.emnlp-main.522/.

APPENDIX CONTENTS

A Foundations and the exact latent representation **13**
A.1 Basic information quantities . . . 13
A.2 Proof of Proposition 1 . . . 13
A.3 Why resetting the prior is a different process . . . 14
A.4 Entropy of form after meaning is fixed . . . 14

B Refinement, semantic decisions, and temperature **15**
B.1 Proof of Proposition 2 . . . 15
B.2 Sequence temperature exposes a form factor . . . 15
B.3 The mode of strings and the mode of meanings . . . 16
B.4 Local temperature has a different normalization . . . 16

C Proof of the binary semantic reversal **17**
C.1 An explicit full-support law . . . 17
C.2 The bad string is the unique global mode . . . 17
C.3 Exact local and global cooled masses . . . 17
C.4 Completing the proof of Theorem 1 . . . 18
C.5 Rates and boundary cases . . . 18

D The sharp path envelope and its critical scale **19**
D.1 A scalar convexity lemma . . . 19
D.2 Proof of Theorem 2 . . . 19
D.3 Deriving the critical scale . . . 20

E Sequence log loss and semantic-error certificates **21**
E.1 The excess log-loss identity . . . 21
E.2 Semantic and conditional form divergence . . . 21
E.3 Proof of Proposition 3 . . . 21
E.4 What a length-independent certificate would require . . . 22

F Proofs for the predictive-memory tradeoff **23**
F.1 The information interface in full . . . 23
F.2 Optimal decoder and actual predictor loss . . . 23
F.3 Converse for arbitrary joint encoders . . . 23
F.4 Attainability and completion of Theorem 3 . . . 24
F.5 Near-optimal predictors and binary Fano . . . 24
F.6 Proof of Corollary 1 . . . 25
F.7 Why the assumptions change the answer . . . 25

G Faithful semantic interfaces **26**
G.1 The cost of approximate semantic masses . . . 26
G.2 Proof of Proposition 4 . . . 26
G.3 An exact autoregressive implementation . . . 27
G.4 What semantic separation does not supply . . . 27

H A charged serialization bound **28**
H.1 One checkpoint and its resources . . . 28
H.2 Proof of both counting bounds . . . 28
H.3 Interpretation without a hidden runtime claim . . . 28

A FOUNDATIONS AND THE EXACT LATENT REPRESENTATION

This appendix separates definitions from results. Source and fixed-horizon output spaces are finite; the generated transcript may have an almost surely finite length with no deterministic upper bound. Contexts may be fixed or drawn from a finite distribution. Public seeds need not be discrete; conditional identities hold almost surely. Conditional laws on zero-probability events may be chosen arbitrarily; a summand with zero weight is omitted. We use $0 \log 0 = 0$ and interpret $p \log(p/0) = +\infty$ for $p > 0$. Results are promptwise unless stated otherwise.

A.1 BASIC INFORMATION QUANTITIES

For finite laws $\mathbb{P}, \mathbb{Q}$, entropy and relative entropy are defined by

$$H(\mathbb{P}) = -\sum_y \mathbb{P}(y) \log \mathbb{P}(y), \quad D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) = \sum_y \mathbb{P}(y) \log \frac{\mathbb{P}(y)}{\mathbb{Q}(y)}. \tag{26}$$

Conditional entropy averages conditional-law entropy. Conditional mutual information is

$$I(A; B \mid R) = H(A \mid R) - H(A \mid B, R). \tag{27}$$

We use nonnegativity of relative entropy, the entropy and relative-entropy chain rules, and the bound $H(H \mid R) \le \log |\mathcal{H}|$. These are standard finite information-theoretic results (Cover & Thomas, 2006, Chapter 2). If the state alphabet depends on R, the bound applies pointwise in R, then averages. Our capacity arguments never equate real-coordinate and bit counts.

A.2 PROOF OF PROPOSITION 1

Proposition 1 (Marginalization does not lose the language law)**.** *The faithful generator $Z \sim \pi(\cdot \mid x)$, $Y \sim r(\cdot \mid x, Z)$ induces the same observable law as the following mixture:*

$$\mathbb{P}(y \mid x) = \sum_z \pi(z \mid x) r(y \mid x, z). \tag{3}$$

At every positive-probability prefix u, the posterior-weighted next-token law is

$$\mathbb{P}(a \mid x, u) = \sum_z \mathbb{P}(z \mid x, u) r(a \mid x, u, z). \tag{4}$$

Conversely, any law $\mathbb{P}$ and any semantic map g_x admit this faithful latent representation.

Proof. Fix x. Suppose first that $Z \sim \pi(\cdot \mid x)$ and $Y \sim r(\cdot \mid x, Z)$. The law of total probability gives (3), whether or not the generator is faithful. Faithfulness says that $r(y \mid x, z) > 0$ implies $g_x(y) = z$, so the generated latent label also equals the output's semantic label.

For a prefix u of length $t - 1$, let

$$r(u \mid x, z) = \sum_{y : y_{<t} = u} r(y \mid x, z).$$

If $\mathbb{P}(u \mid x) > 0$, Bayes' rule gives

$$\mathbb{P}(z \mid x, u) = \frac{\pi(z \mid x) r(u \mid x, z)}{\mathbb{P}(u \mid x)}. \tag{28}$$

For a term with $r(u \mid x, z) > 0$, define

$$r(a \mid x, u, z) = \frac{r(ua \mid x, z)}{r(u \mid x, z)}.$$

Substitution yields

$$\begin{aligned} \sum_z \mathbb{P}(z \mid x, u) r(a \mid x, u, z) &= \frac{\sum_z \pi(z \mid x) r(ua \mid x, z)}{\mathbb{P}(u \mid x)} \\ &= \frac{\mathbb{P}(ua \mid x)}{\mathbb{P}(u \mid x)} = \mathbb{P}(a \mid x, u). \end{aligned} \tag{29}$$

Zero-posterior-weight terms vanish. Multiplying exact next-token conditionals telescopes to the complete-string probability, so posterior-weighted marginalization preserves the observed law.

Conversely, let $\mathbb{P}$ and g_x be given. Set $\pi(z \mid x) = \mathbb{P}(g_x(Y) = z \mid x)$. For $\pi(z \mid x) > 0$, define

$$r(y \mid x, z) = \frac{\mathbb{P}(y \mid x)\mathbf{1}\{g_x(y) = z\}}{\pi(z \mid x)}. \tag{30}$$

Normalized on $\mathcal{Y}_z$, it satisfies (3); discarding zero-mass labels represents every finite law. □

A.3 Why resetting the prior is a different process

A fair latent bit Z with $Y_1 = Y_2 = Z$ gives 00 and 11 mass $1/2$ each. After the first token, the posterior on Z is a point mass. Equation (4) repeats that token with probability one.

Instead extend the component rule to all prefixes by $r(a \mid u, z) = \mathbf{1}\{a = z\}$ and use the original fair prior at each position. Both tokens are now independently fair. The resulting law assigns mass $1/4$ to 01 and to 10, which were impossible under the original generator. This failure comes from replacing the updated posterior by the fixed prior, not from exact marginalization itself.

The example also explains why the existence of a latent representation does not identify a unique internal meaning variable. A different task partition of the same strings gives a different faithful representation. Additional structural or grounding assumptions are required to select a particular semantic interpretation. Proposition 1 is not an identifiability theorem.

A.4 Entropy of form after meaning is fixed

As $Z = g_x(Y)$ is deterministic, $H(Z \mid Y, x) = 0$. The two chain-rule orders give

$$\begin{aligned} H(Y, Z \mid x) &= H(Y \mid x) + H(Z \mid Y, x) = H(Y \mid x), \\ H(Y, Z \mid x) &= H(Z \mid x) + H(Y \mid Z, x). \end{aligned} \tag{31}$$

This proves (5). It is a standard chain-rule specialization, not a new entropy identity. For example, if a meaning has M uniform expressions, its conditional form entropy is $\log M$. If one expression carries almost all its mass, the same cardinality M can coexist with arbitrarily small conditional entropy. Expression counts alone establish no resource allocation result.

B Refinement, semantic decisions, and temperature

B.1 Proof of Proposition 2

Proposition 2 (The only universally refinement-compatible pointwise rule)**.** *Transforming and aggregating commute for every finite law $\mathbb{P}$ and every semantic partition if and only if $w(p) = cp$ for some $c > 0$. Thus ordinary sampling is the only universally compatible rule within this class.*

For a map g, write $g_*\mathbb{P}$ for the aggregated law, so $(g_*\mathbb{P})(z) = \sum_{y:g(y)=z} \mathbb{P}(y)$. The requirement is

$$g_*(T_w\mathbb{P}) = T_w(g_*\mathbb{P}) \tag{32}$$

for every finite distribution and every partition. It includes merging any two forms of a meaning, the inverse of a refinement. The same scalar function w is used on every outcome space.

Proof. If $w(p) = cp$ with $c > 0$, then $T_w\mathbb{P} = \mathbb{P}$, so (32) follows immediately.

For necessity, choose $a, b > 0$ with $a + b < 1$ and set $c_0 = 1 - a - b > 0$. Merge the first two outcomes of (a, b, c_0). The unmerged outcome must have equal probability under both operations:

$$\frac{w(c_0)}{w(a) + w(b) + w(c_0)} = \frac{w(c_0)}{w(a + b) + w(c_0)}.$$

All denominators and $w(c_0)$ are positive. Hence

$$w(a + b) = w(a) + w(b) \;\; (a, b > 0,\; a + b < 1). \tag{33}$$

Continuity extends it to $a + b = 1$; $w(0) = 0$ covers zero arguments. Repeated addition shows $w(1/n) = w(1)/n$ and $w(k/n) = kw(1)/n$ for integers $0 \le k \le n$. Approximate any $p \in [0, 1]$ by rationals and use continuity. Thus $w(p) = w(1)p$. By assumption, $w(1)$ is positive. □

Endpoint continuity matters. If it is removed, one can keep $w(p) = cp$ on $[0, 1)$ and change $w(1)$ to any positive value. The normalized transform remains the identity, because an outcome of probability one occurs only in a point-mass distribution. Our assumptions exclude this endpoint freedom.

The proposition does not cover arbitrary functions of the entire distribution, semantic-aware decoders, or decisions optimized for a supplied task loss. It only characterizes normalized, coordinate-wise probability transformations that commute with *every* aggregation. It follows from additivity.

B.2 Sequence temperature exposes a form factor

For $\beta > 0$, define the sequence law

$$\mathbb{Q}_\beta^{\mathrm{seq}}(y) = \frac{\mathbb{P}(y)^\beta}{\sum_v \mathbb{P}(v)^\beta}.$$

For $y \in \mathcal{Y}_z$, $\mathbb{P}(y) = \pi(z)r(y \mid z)$. Consequently

$$\begin{aligned}\sum_{y \in \mathcal{Y}_z} \mathbb{P}(y)^\beta &= \pi(z)^\beta \sum_{y \in \mathcal{Y}_z} r(y \mid z)^\beta \\ &= \pi(z)^\beta A_\beta(z).\end{aligned} \tag{34}$$

Normalizing across z proves (7). If r is uniform over n_z forms, each summand is $n_z^{-\beta}$ and $A_\beta(z) = n_z^{1-\beta}$. For $\beta > 1$, every conditional law on these available forms satisfies the bounds

$$n_z^{1-\beta} \le A_\beta(z) \le 1. \tag{35}$$

The lower bound follows from Jensen's inequality for $v \mapsto v^\beta$; the upper bound follows from $r(y \mid z)^\beta \le r(y \mid z)$. Uniformity attains the lower bound, and a point mass attains the upper bound. If all n_z probabilities must be positive and $n_z > 1$, the upper bound is an unattained supremum.

Thus form counts bound possible pressure, not its actual value. Equation (7) permits reversals in either direction as form laws change, implying no intrinsic correctness advantage for either class.

B.3 THE MODE OF STRINGS AND THE MODE OF MEANINGS

Draw a reference label Z^* from π and emit a fixed y. Its semantic zero–one risk is

$$\mathbb{P}(g(y) \neq Z^*) = 1 - \pi(g(y)).$$

The Bayes action chooses a class maximizing π and any of its expressions. String MAP maximizes

$$\pi(z) \max_{y \in \mathcal{Y}_z} r(y \mid z). \tag{36}$$

This is ordinary Bayesian decision theory specialized to semantic classes, consistent with the surface form competition literature (Holtzman et al., 2021; Wiegreffe et al., 2023).

For a direct example, give class 0 mass $1 - d$ over M uniform expressions and class 1 mass d on one expression, where $0 < d < 1/2$. If $M > (1 - d)/d$, string MAP chooses class 1. Its semantic risk is $1 - d$, whereas the semantic Bayes risk is d. The excess risk $1 - 2d$ can approach one. This classical form-splitting argument is not the new binary sequential construction of Theorem 1.

B.4 LOCAL TEMPERATURE HAS A DIFFERENT NORMALIZATION

Multiplying (8) along a complete path gives

$$\mathbb{Q}_\beta^{\mathrm{loc}}(y) = \frac{\mathbb{P}(y)^\beta}{\prod_{t=1}^{T} Z_\beta(y_{<t})}, \quad Z_\beta(u) = \sum_{b \in \mathcal{V}} \mathbb{P}(b \mid u)^\beta. \tag{37}$$

This follows directly from the token chain rule and is the local/global distinction studied by Kempton & Burrell (2025). The product in the denominator generally varies with y. It can be constant for special laws, in which case local and global temperature agree; they are not invariably different.

For finite $\beta > 0$, both temperature rules preserve $\mathbb{P}$'s support. A path has positive probability exactly when all path conditionals are positive; positive powers preserve this property. Neither rule can create positive probability on a zero-mass event under an exact reference law. Theorem 1 instead amplifies a positive event whose mass tends to zero along a model family.

C PROOF OF THE BINARY SEMANTIC REVERSAL

Theorem 1 (Accurate samples, incorrect confident decoding)**.** *Fix any $\beta > 1$ and choose $a \in (1/\beta, 1)$. For sufficiently large T, set $\eta = T^{-a}$ in the binary model above. As $T \to \infty$,*

$$\mathbb{P}(\mathsf{bad}) \longrightarrow 0, \quad \mathbb{Q}_\beta^{\mathrm{loc}}(\mathsf{bad}) \longrightarrow 1, \quad \mathbb{Q}_\beta^{\mathrm{seq}}(\mathsf{bad}) \longrightarrow 1. \tag{9}$$

For all sufficiently large T, both greedy decoding and exact sequence MAP output bad*. In contrast, sharpening the two semantic masses rather than individual string probabilities gives*

$$\mathbb{Q}_\beta^{\mathrm{sem}}(\mathsf{bad}) := \frac{\mathbb{P}(\mathsf{bad})^\beta}{\mathbb{P}(\mathsf{bad})^\beta + \mathbb{P}(\mathsf{good})^\beta} \longrightarrow 0. \tag{10}$$

C.1 AN EXPLICIT FULL-SUPPORT LAW

Fix $T \geq 1$, $0 < \eta < 1/2$, and $p = 1 - \eta$. Section 3 defines $\mathbb{P}_{T,\eta}$; when $y = 1^T$,

$$\mathbb{P}_{T,\eta}(y) = p^T. \tag{38}$$

If the first zero of y occurs at position $t \in \{1, \ldots, T\}$, then

$$\mathbb{P}_{T,\eta}(y) = p^{t-1}\eta\, 2^{-(T-t)}. \tag{39}$$

The 2^{T-t} strings have total mass $p^{t-1}\eta$. Every string has positive mass; summing over first-zero positions gives $1 - p^T$, proving normalization. Set $g_T(1^T) = \mathsf{bad}$ and $g_T(y) = \mathsf{good}$ otherwise.

The semantic task is fully specified by the known event "a zero occurs"; no parser must infer the label. Good strings encode one task-relevant outcome, differing only in task-irrelevant tails. This small operational notion does not claim to model all natural-language semantics.

C.2 THE BAD STRING IS THE UNIQUE GLOBAL MODE

Using (39), the likelihood ratio between the bad string and any string with first zero at t is

$$\frac{\mathbb{P}_{T,\eta}(1^T)}{\mathbb{P}_{T,\eta}(y)} = \frac{p}{\eta}(2p)^{T-t} > 1. \tag{40}$$

Both factors are strictly greater than or equal to one, with $p/\eta > 1$, so the inequality is strict even when $t = T$. Thus 1^T is the unique complete-string MAP. Greedy decoding also chooses 1 at each prefix it reaches, because $p > 1/2$. There is no tie-breaking dependence.

Notice the difference from a search failure. A wider beam cannot repair a decision whose exact global objective already prefers the wrong semantic class. Conversely, the result says nothing about an algorithm that aggregates semantic classes rather than maximizing individual strings.

C.3 EXACT LOCAL AND GLOBAL COOLED MASSES

Local temperature gives each all-one-path step conditional probability $p^\beta/(p^\beta + \eta^\beta)$. Hence

$$\mathbb{Q}_\beta^{\mathrm{loc}}(\mathsf{bad}) = \left[1 + \left(\frac{\eta}{p}\right)^\beta\right]^{-T}. \tag{41}$$

The fair conditional after a zero stays fair under cooling. It does not affect the event in (41).

Group probability powers by first zero to obtain the ratio of good to bad powered mass:

$$\begin{aligned} R_{T,\eta,\beta} &:= \frac{\sum_{y \neq 1^T} \mathbb{P}_{T,\eta}(y)^\beta}{p^{\beta T}} \\ &= \frac{\sum_{t=1}^{T} 2^{T-t}\left[p^{t-1}\eta\, 2^{-(T-t)}\right]^\beta}{p^{\beta T}} \\ &= \left(\frac{\eta}{p}\right)^\beta \sum_{j=0}^{T-1} \left(2^{1-\beta}p^{-\beta}\right)^j. \end{aligned} \tag{42}$$

Therefore

$$\mathbb{Q}_\beta^{\mathrm{seq}}(\mathsf{bad}) = \frac{1}{1 + R_{T,\eta,\beta}}. \tag{43}$$

At geometric ratio one, (42) gives sum T without division by a vanishing denominator.

C.4 Completing the proof of Theorem 1

Proof. Fix $\beta > 1$, $a \in (1/\beta, 1)$, and $\eta = T^{-a}$. For sufficiently large T, $0 < \eta < 1/2$. Since $T\eta = T^{1-a} \to \infty$, the reference probability of the all-one path vanishes:

$$\mathbb{P}_{T,\eta}(\mathsf{bad}) = (1-\eta)^T \le \exp(-T\eta) \longrightarrow 0.$$

Let $x_T = (\eta/p)^\beta$. Then $Tx_T = T^{1-a\beta}p^{-\beta} \to 0$. The conditional probability of leaving the all-one path under local cooling is $x_T/(1+x_T)$. A union bound over its T opportunities gives

$$0 \le 1 - \mathbb{Q}_\beta^{\mathrm{loc}}(\mathsf{bad}) \le \frac{Tx_T}{1+x_T} \le Tx_T \longrightarrow 0.$$

This proves the first two limits.

For global cooling, let $c_\beta = 2^{1-\beta} < 1$. Since $p \to 1$, the ratio $c_\beta p^{-\beta}$ tends to c_β. It is thus bounded by a fixed number $r < 1$ for all sufficiently large T. Equation (42) yields

$$0 \le R_{T,\eta,\beta} \le \left(\frac{\eta}{p}\right)^\beta \frac{1}{1-r} \longrightarrow 0.$$

Equation (43) now gives the third limit.

The MAP and greedy claims follow from (40). With $d_T = p^T \to 0$, semantic-first cooling gives

$$\mathbb{Q}_\beta^{\mathrm{sem}}(\mathsf{bad}) = \frac{d_T^\beta}{d_T^\beta + (1-d_T)^\beta} \le \left(\frac{d_T}{1-d_T}\right)^\beta \longrightarrow 0.$$

This completes the proof. □

C.5 Rates and boundary cases

The proof gives sharper rates without trained-model assumptions. In the same regime,

$$1 - \mathbb{Q}_\beta^{\mathrm{loc}}(\mathsf{bad}) \sim T^{1-a\beta}, \quad 1 - \mathbb{Q}_\beta^{\mathrm{seq}}(\mathsf{bad}) \sim \frac{T^{-a\beta}}{1-2^{1-\beta}}, \quad \mathbb{Q}_\beta^{\mathrm{sem}}(\mathsf{bad}) \sim d_T^\beta. \tag{44}$$

For the local rate, use $-\ln \mathbb{Q}^{\mathrm{loc}} = T\ln(1+x_T) \sim Tx_T \to 0$ and $1 - e^{-v} \sim v$. For the global rate, the geometric sum in (42) tends to $1/(1-c_\beta)$; its omitted tail vanishes uniformly once the ratio is at most $r < 1$. The semantic rate follows because its denominator tends to one.

The reference rate is best written as

$$\ln d_T = T\ln(1-T^{-a}) = -T^{1-a}\big(1 + O(T^{-a})\big).$$

This need not imply $d_T \sim \exp(-T^{1-a})$ when $a \le 1/2$. At $a = 1/\beta$, local cooled bad mass tends to e^{-1}, not one. At $a = 1$, reference bad mass tends to e^{-1}, not zero. At $\beta = 1$, all three rules give their unsharpened semantic law. We fix $\beta > 1$ before $T \to \infty$, with no uniform claim as $\beta \to 1^+$.

The family has a two-state recurrent flag, a horizon, and scalar parameter η. Its probabilities vary with T; the logit difference $\ln(p/\eta)$ grows on the order of $\ln T$. This is not a single infinite process with uniformly bounded logits. Every finite member has finite logits and full support.

D THE SHARP PATH ENVELOPE AND ITS CRITICAL SCALE

D.1 A SCALAR CONVEXITY LEMMA

Lemma 1. *For $\beta > 1$, the function*

$$G_\beta(s) = \ln\left(1 + (e^s - 1)^\beta\right)$$

is strictly convex on $[0, \ln 2]$.

Proof. For $0 < s \le \ln 2$, put $v = e^s - 1 \in (0, 1]$. Direct differentiation gives

$$\begin{aligned} G'_\beta(s) &= \frac{\beta(1+v)v^{\beta-1}}{1+v^\beta}, \\ G''_\beta(s) &= \frac{\beta(1+v)v^{\beta-2}\left[(\beta-1) + \beta v - v^\beta\right]}{(1+v^\beta)^2}. \end{aligned} \tag{45}$$

Since $v^\beta \le v$, the bracket is at least $(\beta - 1)(1 + v) > 0$. All other factors are positive. Therefore the derivative is strictly increasing on the interior. At zero, G_β is continuous, $G_\beta(0) = 0$, and $G'_\beta(0+) = 0$. This extends strict convexity to the closed interval. For $1 < \beta < 2$, the second derivative diverges at zero, but no finite second derivative at that endpoint is required. □

D.2 PROOF OF THEOREM 2

Theorem 2 (Sharp amplification on a locally preferred path)**.** *For a binary law, let $p_t = \mathbb{P}(1 \mid 1^{t-1}) \in [1/2, 1]$ and $d = \mathbb{P}(1^T) = \prod_t p_t$. For $\beta > 1$, the cooled path mass satisfies*

$$\mathbb{Q}_\beta^{\mathrm{loc}}(1^T) \le B_{\beta,T}(d) := \left[1 + \left(d^{-1/T} - 1\right)^\beta\right]^{-T}, \quad d \in [2^{-T}, 1]. \tag{11}$$

Equality holds exactly when $p_1 = \cdots = p_T = d^{1/T}$, and is attained with full support for each $d < 1$ in this interval. The tied fair path is allowed, but full support excludes $d = 1$.

Proof. Set $s_t = -\ln p_t$, so $s_t \in [0, \ln 2]$ and $\sum_t s_t = -\ln d$. Cooling chooses 1 with probability

$$\frac{p_t^\beta}{p_t^\beta + (1-p_t)^\beta} = \exp\big(-G_\beta(s_t)\big).$$

Multiplication and Jensen's inequality yield

$$\begin{aligned} -\ln \mathbb{Q}_\beta^{\mathrm{loc}}(1^T) &= \sum_{t=1}^{T} G_\beta(s_t) \\ &\ge T G_\beta\left(\frac{-\ln d}{T}\right) \\ &= T \ln\left[1 + \left(d^{-1/T} - 1\right)^\beta\right]. \end{aligned} \tag{46}$$

Exponentiation gives (11). Equality holds iff all s_t, hence p_t, coincide; this includes $T = 1$.

For attainability, take $p_t = d^{1/T}$ on the all-one path and fair conditional probabilities after the first zero. If $d \in [2^{-T}, 1)$, both choices at every reachable prefix have positive probability, giving full support and attaining the bound. If $d = 1$, every path conditional equals one and the bound equals one; this endpoint is attained only if full support is not required. No full-support model has $d = 1$. The value $B_{\beta,T}(1) = 1$ is approached by full-support models as $d \to 1^-$. At $d = 2^{-T}$, the fair model attains the bound 2^{-T}; Theorem 1 makes no strict-MAP claim at this tied endpoint. □

Off-path conditionals enter the proof only to ensure a valid model. The theorem controls a specified path that is locally preferred at every step. It does not characterize all semantic events or paths containing a less likely token. Such generalizations require a different optimization problem.

D.3 Deriving the critical scale

Fix $\beta > 1$ and let $L_T = -\ln d_T > 0$, with $L_T = o(T)$ and $u_T = L_T/T \to 0^+$. The expansions $e^{u_T} - 1 \sim u_T$ and $\ln(1+v) \sim v$ as $v \to 0^+$ give the leading order of the finite-length envelope:

$$\begin{aligned} -\ln B_{\beta,T}(d_T) &= T \ln\left(1 + (e^{u_T} - 1)^\beta\right) \\ &\sim T u_T^\beta = \frac{L_T^\beta}{T^{\beta-1}}. \end{aligned} \tag{47}$$

This proves (12). More explicitly:

- If $L_T = o(T^{1-1/\beta})$, the envelope tends to one.
- If $L_T/T^{1-1/\beta} \to c \in (0, \infty)$, it tends to $\exp(-c^\beta)$.
- If $T^{1-1/\beta} = o(L_T)$ while $L_T = o(T)$, it tends to zero.

The first regime is compatible with $L_T \to \infty$, meaning $d_T \to 0$. Thus a vanishing reference path probability gives no horizon-independent guarantee after fixed local cooling. Equal path conditionals attain the envelope at every finite horizon, so no loose asymptotic estimate is involved.

For Theorem 1's family, $L_T \sim T^{1-a}$ and $1 - a < 1 - 1/\beta$ exactly when $a > 1/\beta$. The critical-scale result recovers this threshold and explains why strict inequality is necessary.

E SEQUENCE LOG LOSS AND SEMANTIC-ERROR CERTIFICATES

E.1 THE EXCESS LOG-LOSS IDENTITY

For finite $D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})$, the expected excess log loss is

$$\begin{aligned}\mathbb{E}_{Y\sim\mathbb{P}(\cdot)}[-\log\mathbb{Q}(Y)+\log\mathbb{P}(Y)] &= \sum_y \mathbb{P}(y)\log\frac{\mathbb{P}(y)}{\mathbb{Q}(y)}\\ &= D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}).\end{aligned} \tag{48}$$

Using the chain factorization of $\mathbb{P}$ and $\mathbb{Q}$, this also equals

$$\sum_{t=1}^{T}\mathbb{E}_{Y_{<t}\sim\mathbb{P}(\cdot)}D_{\mathrm{KL}}(\mathbb{P}(\cdot\mid Y_{<t})\|\mathbb{Q}(\cdot\mid Y_{<t}))\,. \tag{49}$$

Thus (13) is mean *teacher-forced* excess log loss over the reference prefixes. It is neither a uniform error bound for every prefix nor an expectation over the model's own generated prefix distribution. These standard identities (Cover & Thomas, 2006) specify which quantity is normalized by T.

E.2 SEMANTIC AND CONDITIONAL FORM DIVERGENCE

For each label with $\mathbb{P}_Z(z)>0$, finite sequence KL implies $\mathbb{Q}_Z(z)>0$. For $y\in\mathcal{Y}_z$ with $\mathbb{P}(y)>0$,

$$\log\frac{\mathbb{P}(y)}{\mathbb{Q}(y)}=\log\frac{\mathbb{P}_Z(z)}{\mathbb{Q}_Z(z)}+\log\frac{\mathbb{P}(y\mid z)}{\mathbb{Q}(y\mid z)}.$$

Multiply by $\mathbb{P}(y)$, sum within each class, and then sum over z. This gives (14). If $\mathbb{Q}_Z(z)=0<\mathbb{P}_Z(z)$ or a required conditional assigns zero probability to a reference outcome, both sides have the appropriate value $+\infty$. Zero-weight classes are omitted, avoiding undefined conditional laws. Nonnegativity of the conditional terms implies $D_{\mathrm{KL}}(\mathbb{P}_Z\|\mathbb{Q}_Z)\le D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})$.

This proof uses the same fixed semantic map for the two laws. Changing the semantic map between models would compare different tasks and would not justify this data-processing conclusion.

E.3 PROOF OF PROPOSITION 3

Proposition 3 (A sharp semantic-error certificate)**.** *Let E be the event that the output has an incorrect semantic label. If $\mathbb{P}(E)=0$ and $\epsilon_T(\mathbb{P},\mathbb{Q})\le\epsilon$, then ordinary sampling from $\mathbb{Q}$ satisfies*

$$\mathbb{Q}(E)\le 1-2^{-T\epsilon}. \tag{15}$$

For every finite T and $\epsilon\ge 0$, the bound is attained whenever a correct and an incorrect output are available. There also exist binary examples with $\epsilon_T\to 0$ but $\mathbb{Q}(E)\to 1$ as $T\to\infty$.

Proof. Let $e=\mathbb{Q}(E)$. The condition $\mathbb{P}(E)=0$ puts reference mass on E^c. Finite KL excludes $e=1$. For $e<1$, set $\mathbb{Q}_0=\mathbb{Q}(\cdot\mid E^c)$, so $\mathbb{Q}(y)=(1-e)\mathbb{Q}_0(y)$ on the reference support. Hence

$$\begin{aligned}D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) &= -\log(1-e)+D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}_0)\\ &\ge -\log(1-e).\end{aligned} \tag{50}$$

Combining with $D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})\le T\epsilon$ gives $e\le 1-2^{-T\epsilon}$.

For sharpness, take any law $\mathbb{P}$ supported on E^c and any law $\mathbb{S}$ supported on E. Set $e=1-2^{-T\epsilon}$ and $\mathbb{Q}=(1-e)\mathbb{P}+e\mathbb{S}$. For every y with $\mathbb{P}(y)>0$, $\mathbb{Q}(y)=(1-e)\mathbb{P}(y)$. Thus $D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})=-\log(1-e)=T\epsilon$, and $\mathbb{Q}(E)=e$. This includes $\epsilon=0$, where $\mathbb{Q}=\mathbb{P}$. The assumed availability of both correct and incorrect output strings ensures that these two component laws exist.

For the varying-horizon claim, let the first $T-1$ tokens be independent fair bits under both laws. The reference has final token zero. The learned law has final token one with probability

$$e_T=1-2^{-\sqrt{T}},$$

independently of the prefix. Let the semantic map read the final token, marking one incorrect. The first $T-1$ conditional KL terms vanish; the last is $-\log(1-e_T) = \sqrt{T}$. Hence

$$\epsilon_T(\mathbb{P}, \mathbb{Q}) = \frac{1}{\sqrt{T}} \longrightarrow 0, \;\; \mathbb{Q}(E) = e_T \longrightarrow 1.$$

The reference is truthful for every prefix, so no error in the reference data is used. □

E.4 What a length-independent certificate would require

The last construction does not have vanishing sequence KL. For fixed T, the upper bound in (15) tends to zero as $\epsilon \to 0$. For varying T, requiring $T\epsilon_T \to 0$ is sufficient. A separate semantic loss can also certify the task directly rather than averaging over irrelevant positions. Our result rules out a universal guarantee based only on $\epsilon_T \to 0$ without accounting for horizon.

The error event is task-specific. In the last-bit construction, the other bits are form by definition. Identifying task-relevant subsets in natural language may be difficult. This exactly refutes the proposed implication but does not claim that every long explanation dilutes its own factual reliability.

F PROOFS FOR THE PREDICTIVE-MEMORY TRADEOFF

Theorem 3 (The exact cost of semantic retention). *Let $1 \le C \le m$ be an integer. The optimum over all admissible encoders and the semantic-retention cost for any such encoder obey*

$$\mathcal{L}_* = \frac{m+\lambda-C}{m+\lambda}, \quad \mathcal{L}(H) - \mathcal{L}_* \ge \frac{1-\lambda}{m+\lambda} I(Z; H \mid R). \tag{18}$$

With source- and query-independent public time sharing, the bound is attained at every retention level in $[0,1]$. Every exact optimum has $I(Z; H \mid R) = 0$ and semantic Bayes error $1/2$. Protecting the entire semantic bit costs exactly $(1-\lambda)/(m+\lambda)$ in loss, at unchanged capacity.

F.1 THE INFORMATION INTERFACE IN FULL

Take independent (U, Z): $U = (U_1, \dots, U_m) \sim \mathrm{Unif}(\{0,1\}^m)$, $Z \sim \mathrm{Unif}(\{0,1\})$. Require

$$\mathbb{P}_{U,Z,R,J,H} = \mathbb{P}_U \mathbb{P}_Z \mathbb{P}_R \mathbb{P}_J \mathbb{P}_{H|U,Z,R}. \tag{51}$$

Queries follow (16). For every public seed value, the encoded state has at most 2^C values. The encoder may use private random bits, with no side information about J. The decoder has only (H, R, J) and independent fresh randomness. The target A is U_J if $J \ge 1$ and Z if $J = 0$.

Equation (51) requires both source independence of the public seed and query independence conditional on that seed. The latter cannot be dropped. If the seed revealed the query, an encoder could store just the requested bit, changing the optimum. Retaining the erased source in a prompt, cache, side channel, or public transcript changes the capacity and invalidates the stated interface.

Every loss evaluation uses one delayed binary query after one encoding. No successive coordinates are teacher-forced: that could reveal earlier answers, defining a different problem.

F.2 OPTIMAL DECODER AND ACTUAL PREDICTOR LOSS

For any fixed encoder and any probabilistic decoder q, conditional expected log loss decomposes as

$$\begin{aligned}\mathcal{L}(q, H) &= H(A \mid H, R, J) \\ &\quad + \mathbb{E}_{(H,R,J)\sim\mathbb{P}_{H,R,J}(\cdot)} D_{\mathrm{KL}}(\mathbb{P}(A \in \cdot \mid H, R, J) \| q(\cdot \mid H, R, J)) .\end{aligned} \tag{52}$$

This follows by adding and subtracting the log of the true conditional probability, as in (48). The posterior decoder attains equality with the first term. The independence in (51) implies

$$H(A \mid H, R, J) = \frac{\sum_i H(U_i \mid H, R) + \lambda H(Z \mid H, R)}{m+\lambda},$$

which proves (17); posterior-decoder lower bounds therefore hold for every approximate decoder.

F.3 CONVERSE FOR ARBITRARY JOINT ENCODERS

Set $I_Z = I(Z; H \mid R)$ and $M_0 = m + \lambda$. Independence gives $H(U_i \mid R) = H(Z \mid R) = 1$, so

$$M_0 \mathcal{L}(H) = M_0 - \sum_i I(U_i; H \mid R) - \lambda I_Z. \tag{53}$$

Information about independent source bits cannot exceed the encoded state's capacity:

$$\begin{aligned}\sum_i I(U_i; H \mid R) &\le I(U; H \mid R), \\ I(U; H \mid R) + I_Z &\le I(U, Z; H \mid R), \\ I(U, Z; H \mid R) &\le H(H \mid R) \le C.\end{aligned} \tag{54}$$

First, apply conditional entropy subadditivity to the right side minus the left. For the second, the exact difference is $I(U; Z \mid H, R)$ since U and Z are independent given R. The last uses the finite state count. These bounds allow encodings that mix coordinates or use private randomness.

Combining (53) and (54),

$$\sum_i I(U_i; H \mid R) + \lambda I_Z \le I(U; H \mid R) + I_Z - (1-\lambda) I_Z$$
$$\le C - (1-\lambda) I_Z, \tag{55}$$

which gives

$$\mathcal{L}(H) \ge \frac{M_0 - C}{M_0} + \frac{1-\lambda}{M_0} I_Z. \tag{56}$$

No coordinate-wise restriction on H has been imposed.

An exact remainder identity makes all possible slack explicit:

$$M_0 \left[\mathcal{L}(H) - \frac{M_0 - C}{M_0} \right] = (1-\lambda) I_Z + [C - I(U, Z; H \mid R)]$$
$$+ I(U; Z \mid H, R)$$
$$+ \sum_i H(U_i \mid H, R) - H(U \mid H, R). \tag{57}$$

Every term on the right is nonnegative. It accounts for unused capacity, dependence induced between the independent sources by observing the code, and dependence among form bits given the code. A synergistic code cannot evade the bound by making each bit individually uninformative.

F.4 Attainability and completion of Theorem 3

Proof. Store any C form bits as $H = (U_1, \ldots, U_C)$, with a constant public seed. The other $m - C$ form queries and the semantic bit have conditional entropy one; stored queries have zero. This gives

$$\mathcal{L}_* = \frac{m - C + \lambda}{m + \lambda},$$

attaining the lower bound in (56) with $I_Z = 0$. This proves the optimum and (18).

Storing $H = (Z, U_1, \ldots, U_{C-1})$ leaves no semantic entropy and $m - C + 1$ unknown form bits:

$$\mathcal{L}(H) = \frac{m - C + 1}{m + \lambda} = \mathcal{L}_* + \frac{1-\lambda}{m+\lambda}, \quad I_Z = 1.$$

The $I_Z = 1$ converse proves that this C-bit code has minimum loss cost for perfect recovery.

For any $r \in [0, 1]$, choose an independent public Bernoulli seed with $\mathbb{P}(R = 1) = r$. Use the second code when $R = 1$ and the first code otherwise. The decoder knows which interpretation to use, but the public seed contains no source or query information. Conditional mutual information and expected loss average across the two branches, giving $I_Z = r$ and $\mathcal{L}(H) - \mathcal{L}_* = (1-\lambda) r/(m+\lambda)$. Hence the entire information–loss line is attainable when this independent time sharing is allowed.

At any optimum $\mathcal{L}_*$, (18) and $\lambda < 1$ force $I_Z = 0$ and $H(Z \mid H, R) = 1$. Only a fair bit has entropy one: $\mathbb{P}(Z = 0 \mid H, R) = \mathbb{P}(Z = 1 \mid H, R) = 1/2$ almost surely, giving best error $1/2$. □

Attainability allows independent time sharing. A preassigned seed distribution need not support every mixing weight. The converse and both endpoints hold without public randomness.

F.5 Near-optimal predictors and binary Fano

Suppose the predictor obeys $\mathcal{L}(q, H) \le \mathcal{L}_* + \Delta$. By (52), $\mathcal{L}(H) \le \mathcal{L}(q, H)$. Then (18) gives

$$I_Z \le \min\left\{1, \frac{(m+\lambda)\Delta}{1-\lambda}\right\}.$$

For a posterior mode $\hat{Z}(H, R)$, set $B = \mathbf{1}\{\hat{Z} \ne Z\}$. Given (H, R), B determines the binary source:

$$H(Z \mid H, R) = H(B \mid H, R)$$
$$\le H(B) = h_2(e_Z). \tag{58}$$

Combining this with $H(Z \mid H, R) = 1 - I_Z$ proves (19). This is the binary case of Fano's inequality (Cover & Thomas, 2006). The posterior mode has $e_Z \le 1/2$, so the inequality can be inverted on $[0, 1/2]$ for a numerical lower bound. Attainability at every Δ remains unproved.

F.6 Proof of Corollary 1

Corollary 1 (Generated text cannot restore erased information)**.** *From (H, R) and fresh randomness jointly independent of (U, Z, H, R), generate an almost surely finite transcript W with no length or computation bound. It adds no semantic information beyond the retained state:*

$$H(Z \mid H, R, W) = H(Z \mid H, R). \tag{20}$$

The best semantic error using (H, R, W) is exactly e_Z, the error using (H, R) alone. In particular, every optimal encoder in Theorem 3 remains at semantic error $1/2$ after any such transcript.

Proof. Let the fresh random tape be jointly independent of (U, Z, H, R). Generate W, including its stopping decision, from that tape and (H, R) alone. Thus W and Z are conditionally independent given (H, R), even with no deterministic bound on the almost surely finite length.

For each z, conditional independence gives the posterior identity below. Conditional expectations also cover continuous public seeds without conditioning on positive-probability atoms:

$$\mathbb{E}[\mathbf{1}\{Z = z\} \mid H, R, W] = \mathbb{E}[\mathbf{1}\{Z = z\} \mid H, R] \text{ almost surely.} \tag{59}$$

Averaging posterior entropy proves (20); averaging one minus the largest posterior mass proves equality of best semantic error. Theorem 3 supplies the $1/2$ conclusion. An independent query can join the fresh inputs if their joint independence from (U, Z, H, R) is preserved.

Source independence alone is insufficient. For independent fair bits Z, N, take constant R, $H = Z \oplus N$, and $W = N$. Although W is independent of Z, (H, W) recovers Z. Reusing the encoder's mask violates the fresh-tape assumption; it is not a counterexample to the corollary. □

This is the conditional-independence equality case underlying the data-processing principle (Cover & Thomas, 2006), not a new information inequality. Giving the generator an additional source-dependent observation breaks the premise. Restricting decoder computation can also change the operational value of the transcript even though its posterior information is unchanged.

F.7 Why the assumptions change the answer

At $\lambda = 1$, the two endpoint codes have equal loss. The nonnegative penalty on I_Z vanishes, and semantic forgetting is not forced. At $C = m + 1$, storing all source bits gives zero loss and zero semantic error. At $C = 0$, no source-dependent state is available, so every target remains fair; this case is excluded because perfect semantic recovery at the same capacity is impossible.

If the query were known before encoding, one bit could store the requested answer, independently of m. If the source were visible after encoding, no persistent-state lower bound of this form would apply. These observations are constructive counterexamples to removing the access assumptions, not technical exceptions to an otherwise universal language-model theorem.

The semantic bit's weight matters. When $\lambda > 1$ and $1 \le C \le m + 1$, (54) gives

$$\sum_i I(U_i; H \mid R) + \lambda I_Z \le C + (\lambda - 1) I_Z \le C + \lambda - 1.$$

Storing Z and $C - 1$ form bits attains the resulting optimum. The same architecture and same number of bits can therefore prioritize semantics under a different objective. The source of pressure is capacity plus task weighting, not the mere existence of more possible forms.

G FAITHFUL SEMANTIC INTERFACES

G.1 THE COST OF APPROXIMATE SEMANTIC MASSES

Let $\|\tilde{\pi} - \pi\|_{\mathrm{TV}} = \frac{1}{2}\sum_z |\tilde{\pi}(z) - \pi(z)| \le \varepsilon$. On finite laws, it bounds every event's absolute mass difference, including singletons (Cover & Thomas, 2006). For the two main-text maximizers,

$$\begin{aligned}\pi(z_\star) - \pi(\tilde{z}) &= \pi(z_\star) - \tilde{\pi}(z_\star) + \tilde{\pi}(z_\star) - \tilde{\pi}(\tilde{z}) \\ &\quad + \tilde{\pi}(\tilde{z}) - \pi(\tilde{z}) \\ &\le |\pi(z_\star) - \tilde{\pi}(z_\star)| + |\tilde{\pi}(\tilde{z}) - \pi(\tilde{z})| \\ &\le 2\varepsilon. \end{aligned} \tag{60}$$

The middle difference is nonpositive since $\tilde{z}$ maximizes $\tilde{\pi}$. Hence (21), including arbitrary ties. It bounds decision regret against the supplied π, not error relative to an independently specified truth. Faithful realization preserves the bound: changing expression leaves the chosen label unchanged.

G.2 PROOF OF PROPOSITION 4

Proposition 4 (Separation of decision and realization)**.** *A faithful decoder induces semantic law $\hat{\pi}$, independently of its within-class laws $\mathbb{K}(\cdot \mid x, z)$. More generally, let $\delta, \rho \in [0, 1]$, let $z^*(x)$ be the target, and let planner error be at most δ. Suppose every reachable (x, z) obeys*

$$\mathbb{K}\big(g_x(Y) \neq z \mid x, z\big) \le \rho. \tag{22}$$

The planner and decoder errors then give the following sharp bound on total semantic error:

$$\mathbb{P}\big(g_X(Y) \neq z^*(X)\big) \le \delta + (1 - \delta)\rho. \tag{23}$$

Adding faithful forms introduces no length-dependent term; the bound remains sharp.

Proof. Fix x and suppose $\mathbb{K}$ is faithful. By the law of total probability,

$$\begin{aligned}\mathbb{P}(g_x(Y) = z \mid x) &= \sum_j \hat{\pi}(j \mid x)\mathbb{K}(g_x(Y) = z \mid x, j) \\ &= \sum_j \hat{\pi}(j \mid x)\mathbf{1}\{j = z\} = \hat{\pi}(z \mid x). \end{aligned} \tag{61}$$

Thus any change to $\mathbb{K}$ that preserves faithfulness leaves the semantic law unchanged. A refined output alphabet may be handled by giving its new strings the same label; the argument is identical.

Now allow random prompts and approximate faithfulness. Write $G = \{\hat{Z} = z^*(X)\}$ for a correct plan and $F = \{g_X(Y) \neq \hat{Z}\}$ for realization failure. Output error lies in $G^c \cup (G \cap F)$, and $d = \mathbb{P}(G^c) \le \delta$. The decoder depends only on $(X, \hat{Z})$, so its pointwise guarantee yields

$$\mathbb{P}(F \mid G) \le \rho$$

whenever G has positive probability. It follows that

$$\begin{aligned}\mathbb{P}(g_X(Y) \neq z^*(X)) &\le d + (1 - d)\rho \\ &\le \delta + (1 - \delta)\rho. \end{aligned} \tag{62}$$

The case $\mathbb{P}(G) = 0$ satisfies the bound trivially when the planner-error premise permits it.

For sharpness, use a fixed prompt with target label zero and one available output for each of two labels. Let the planner output label one with probability δ. Decode label one faithfully. On planned label zero, output label one with probability ρ and zero otherwise. The decoder satisfies the uniform failure upper bound ρ, and the total semantic error is exactly $\delta + (1 - \delta)\rho$. □

The minimal probabilistic premise for the refined composition bound is $\mathbb{P}(F \mid G) \le \rho$. A guarantee averaged only over proposed labels can be insufficient if it allows realization failures to concentrate on correct plans. Our stated kernel and pointwise guarantee avoid that hidden dependence. A looser union bound $\delta + \rho$ would not require the same conditioning care, but would miss the sharp coefficient.

Invariance is a faithful-interface property, not a guarantee for every neural decoder. At each reachable prompt, every label the planner uses with positive probability needs an available expression.

G.3 AN EXACT AUTOREGRESSIVE IMPLEMENTATION

Fix a label z with $\pi(z) > 0$. For a prefix u of positive reference probability, define

$$h_z(u) = \mathbb{P}(g_x(Y) = z \mid Y_{1:|u|} = u).$$

A complete string y has $h_z(y) = \mathbf{1}\{g_x(y) = z\}$. At any unfinished prefix, total probability gives

$$h_z(u) = \sum_{a \in \mathcal{V}} \mathbb{P}(a \mid u) h_z(ua). \tag{63}$$

When $h_z(u) > 0$, the probabilities in (24) sum to one and are nonnegative. A transition with positive $\mathbb{K}_z(a \mid u)$ reaches a prefix with positive $h_z(ua)$, so division by zero never occurs on a realized path.

For a complete output in the conditional support, multiplication telescopes:

$$\begin{aligned} \prod_{t=1}^{T} \mathbb{K}_z(y_t \mid y_{<t}) &= \left[\prod_{t=1}^{T} \mathbb{P}(y_t \mid y_{<t})\right] \frac{h_z(y)}{h_z(\varnothing)} \\ &= \frac{\mathbb{P}(y)\mathbf{1}\{g_x(y) = z\}}{\pi(z)}. \end{aligned} \tag{64}$$

The result is exactly $\mathbb{P}(\cdot \mid Z = z)$, which is faithful. Prefixes outside the decoder's reachable support can have arbitrary transitions. This proves (24) and its asserted properties.

In particular, an existing token factorization can realize the protected semantic law without changing the final output alphabet. What changes is access to the continuation values h_z. They are normalized sums over possible completions, not quantities supplied for free by a next-token probability. The construction proves existence, not an efficient algorithm for computing those sums.

G.4 WHAT SEMANTIC SEPARATION DOES NOT SUPPLY

Introducing another random variable alone gives a latent planner no missing world information. If the observed prompt is compatible with several ground truths, the planner still faces that uncertainty. A faithful decoder faithfully expresses its choice, including an incorrect one. The composition bound isolates planning error; it cannot make δ small without additional evidence or reasoning assumptions.

The interface requires no irrevocable commitment to a complete idea at the first token. A system may revise its semantic state or maintain a plan distribution. Any guarantee must specify the state being realized and its update rule. Our proposition concerns a completed planned label and its realization, not a universal temporal ordering of thought, language, and computation in humans or machines.

H A CHARGED SERIALIZATION BOUND

Main-text counting specializes fixed-length lossless coding (Cover & Thomas, 2006, Chapter 5). The interface matters: uncharged transcripts or changed output requirements invalidate the claim.

H.1 ONE CHECKPOINT AND ITS RESOURCES

Let Z range over N semantic states and U over M prescribed form states. All NM pairs are admissible. A checkpoint encoder sees (Z, U) before this source is erased and produces a retained state $H \in \mathcal{H}$, with $|\mathcal{H}| \le 2^C$, and a visible transcript $W \in \mathcal{A}^L$, with $|\mathcal{A}| = A \ge 2$. The transcript has *exactly* L positions, known to both parties. The pair (H, W) may be jointly encoded in any way.

A source-independent public seed may be available; the encoder and decoder may use private randomness. All source-dependent information accessible after erasure must occur in H or W. A clock whose value depends on the source would be another message and must be charged. Zero error means exact recovery for every source pair, with probability one over the permitted randomness.

H.2 PROOF OF BOTH COUNTING BOUNDS

For a fixed public seed, full recovery forbids distinct source pairs from sharing a positive-probability message (h, w). The decoder would otherwise have to recover both pairs with probability one from the same information. Private randomness cannot resolve this contradiction. The NM source pairs thus need disjoint nonempty message supports. At most $2^C A^L$ messages exist, giving

$$NM \le 2^C A^L.$$

For a random public seed, apply this argument almost surely in the seed. Because there are finitely many source pairs, their probability-one correctness events have a common probability-one intersection. Source independence ensures that the seed does not itself identify a source pair.

For semantic-only recovery, fix one form value and vary the semantic state. Message supports for different semantic labels must again be disjoint, so $N \le 2^C A^L$. Different form values within one label may now share a message, which is why the factor M disappears.

Conversely, choose $|\mathcal{H}| = 2^C$ within the budget. If $2^C A^L \ge NM$, inject the source pairs into $\mathcal{H} \times \mathcal{A}^L$ and invert to decode. For semantic-only recovery, inject the N labels and ignore U. These codes attain the bounds by joint encoding before erasure, not necessarily under every online schedule.

For integer $C \ge 0$, the exact minimum transcript lengths under these assumptions are therefore

$$L_{\text{both}} = \max\left\{0, \left\lceil \frac{\log(NM) - C}{\log A} \right\rceil\right\}, \quad L_{\text{meaning}} = \max\left\{0, \left\lceil \frac{\log N - C}{\log A} \right\rceil\right\}. \tag{65}$$

This proves (25), including rounding. Variable-length messages require counting their lengths and termination protocol; the count A^L applies only under the stated fixed-length assumption.

H.3 INTERPRETATION WITHOUT A HIDDEN RUNTIME CLAIM

The bound concerns a *prescribed* form state. If a later renderer may choose a fresh form, it need not recover U, so the semantic-only bound applies. Complete earlier text that remains available is part of W, not free memory. If unlimited hidden source-dependent state is allowed, it must count in C.

The comparison consequently supports a limited statement: enforcing a lossless surface realization at an internal checkpoint can require more communication than preserving meaning alone. It does not imply a general lower bound on the computational time of chain-of-thought reasoning, and it does not say that a latent model avoids the cost of writing its final answer. Runtime claims require a machine model, a specified computation task, and explicit accounting of computation.