A No-Go Result for Absolute Internal Prohibitions on Conceptual Composition in Language Models

Abstract

We formalize and prove a basic impossibility: if a capable language model must reliably behave differently whenever an input entails the conjunction of two abstract concepts A and B (e.g., to refuse or deflect), then the model's internal states must carry nonzero information about the property "A-and-B" (the conjunction). Thus a system cannot both (i) be generally competent about A and B and (ii) *never* internally represent the conjunction across all contexts. The proof follows from the Data Processing Inequality and a Fano-type lower bound. We also show that such internal information must appear at some time step/layer by the chain rule for mutual information. This formalizes the intuition that there is no censorship without recognition.

1 Setup

Let X be an input random variable ranging over natural-language strings, processed by a model that produces a sequence of internal states $H = (H_1, \ldots, H_T)$ and an output Y. Allow internal model randomness R that is independent of X. We assume H and Y are (possibly randomized) measurable functions of (X, R); in particular, for some measurable F and G, [H = F(X,R), Y = G(H,R).]

Definition 1 (Conjunctive property)**.** *Fix two abstract concepts A and B. Let $P \in 0, 1$ be the* conjunctive property *indicating whether the content of X entails (possibly implicitly, via paraphrase or indirection) the conjunction "A and B". We assume P is a measurable function of X; i.e., $P = g(X)$ for some g.*

We will use standard information-theoretic notation: $I(U; V)$ denotes mutual information between random variables U and V, and $H(\cdot)$ denotes entropy. We write $d_{\mathrm{TV}}(\mu, \nu)$ for total variation distance between distributions μ and ν.

Definition 2 (Never-conceive requirement)**.** *We say the model* never conceives *the conjunction if [I(H;P) = 0.] A stronger, layerwise version demands $I(H_t; P) = 0$ for all t.*

Definition 3 (Reliable restriction requirement)**.** *We say the model* reliably restricts *the conjunction if there exists $\delta > 0$ such that the conditional output laws differ by at least δ in total variation: [$d_{\mathrm{TV}}\big(\mathcal{L}(Y \mid P{=}0), ; \mathcal{L}(Y \mid P{=}1)\big) \geq \delta$.]Equivalently, Yisnotindependentof P, i.e., I(Y;P) ¿ 0.*

Intuitively, reliable restriction covers behaviors such as consistent refusal/deflection when $P = 1$ and normal answering when $P = 0$.

2 Key lemmas

Lemma 1 (Markov structure and DPI)**.** *Under the setup above, the random variables form a Markov chain $P \to X \to H \to Y$. Consequently, by the Data Processing Inequality (DPI), [I(P;Y)* $\le I(P;H) \le I(P;X)$.]

Proof. By definition $P = g(X)$, so $P \to X$ is deterministic; $H = F(X,R)$, so $X \to H$; and $Y = G(H,R)$, so $H \to Y$. Moreover, R is independent of X, hence of P. Thus (P,X,H,Y) is a Markov chain $P \to X \to H \to Y$. DPI then gives $I(P;Y) \le I(P;H) \le I(P;X)$. □

Lemma 2 (Independence characterization)**.** $I(Y;P) = 0$ *if and only if* $\mathcal{L}(Y \mid P{=}0) = \mathcal{L}(Y \mid P{=}1)$. *In particular, if* $d_{\mathrm{TV}}\big(\mathcal{L}(Y \mid P{=}0), ; \mathcal{L}(Y \mid P{=}1)\big) > 0$*, then* $I(Y;P) > 0$.

Proof. $I(Y;P) = 0$ if and only if Y and P are independent, which holds iff their conditional laws are equal almost surely. If the total variation distance between the two conditionals is positive, they differ on a measurable set, so independence fails and $I(Y;P) > 0$. □

Lemma 3 (Layerwise chain rule)**.** *Let* $H = (H_1, \dots, H_T)$*. Then [I(P;H) ;=;* $\sum_{t=1}^{T} I\big(P; H_t, \big|, H_1, \dots, H_{t-1}\big), .$]*Inpa*
$H_t \mid H_{<t}) > 0$.

Proof. This is the chain rule for mutual information. □

3 Main impossibility theorem

Theorem 1 (No-go for absolute internal prohibition)**.** *Suppose the model satisfies the reliable restriction requirement, i.e., there exists* $\delta > 0$ *with [* $d_{\mathrm{TV}}\big(\mathcal{L}(Y \mid P{=}0), ; \mathcal{L}(Y \mid P{=}1)\big) \ge \delta$.]$Thenthenever-$
$conceiverequirementisimpossible$: *necessarilyI(H;P) ¿ 0.Moreover,thereexistsatleastonetimesteptsuchthatI(P*
$H_t \mid H_{<t}) > 0$.

Proof. By Lemma 2, the total variation gap implies $I(Y;P) > 0$. By Lemma 1 (DPI), $I(Y;P) \le I(H;P)$, hence $I(H;P) > 0$, contradicting $I(H;P) = 0$. The layerwise statement follows from Lemma 3: if $I(P;H) > 0$, at least one conditional term must be positive. □

Corollary 1 (No censorship without recognition)**.** *Any system that reliably treats inputs with* $P = 1$ *differently from* $P = 0$ *must* internally represent *information about* P *at some point in its computation. In particular, a generally capable model that can discuss A and B and also reliably suppress their conjunction cannot avoid "conceiving" the conjunction internally.*

4 A quantitative lower bound (Fano-type)

The next result lower-bounds how much information about P must flow to the output if the model distinguishes P with small error.

Theorem 2 (Fano-style lower bound)**.** *Assume* P *is non-degenerate with entropy* $H(P) > 0$. *Suppose there exists a decision rule* $\hat{P} = \hat{P}(Y)$ *such that* $\Pr[\hat{P} \ne P] \le \varepsilon < \frac{1}{2}$. *Then [I(Y;P)* $;\ge; H(P) - h(\varepsilon),]$*whereh*$(\cdot)$ *is the binary entropy function. Consequently, by DPI, [I(H;P)* $;\ge ; I(Y;P); \ge; H(P) - h(\varepsilon); >; 0.]$

Proof. Fano's inequality gives $H(P \mid Y) \le h(\varepsilon)$ for binary P, hence $I(Y;P) = H(P) - H(P \mid Y) \ge H(P) - h(\varepsilon)$. The DPI step follows from Lemma 1. □

Interpretation. If the system's behavior distinguishes P from $\neg P$ with error below chance, then strictly positive information about P must appear at the output, hence also in the internal state by DPI. This quantifies the qualitative no-go of Theorem 1.

5 On "arbitrary restrictions" over open vocabularies

Let $\mathcal{C}$ be a large family of abstract concepts (synonyms, paraphrases, metonymic variants). For each ordered pair $(A, B) \in \mathcal{C} \times \mathcal{C}$, let $P_{A,B}$ denote the property "input entails A and B". Requiring a generally capable model to:

- competently discuss each concept in $\mathcal{C}$, and
- for an *arbitrary* specified subset $\mathcal{S} \subset \mathcal{C} \times \mathcal{C}$, reliably restrict any $P_{A,B}$ with $(A, B) \in \mathcal{S}$,

forces the model to internally recognize each $P_{A,B}$ it treats differently (by Theorem 1), across all paraphrases/contexts. In open vocabularies, $\mathcal{S}$ may be combinatorially large and semantically fuzzy, so the requirement that *no* internal representation of these conjunctions ever appears is incompatible with reliability and general competence.

6 Conclusion

Theorems 1 and 2 formalize an intuitive constraint: to reliably restrict a semantic relation, a model must internally *recognize* that relation. Therefore, a generally capable language model cannot both explain abstract concepts A and B and *never* internally represent their conjunction in any context while also reliably treating that conjunction differently. The achievable target is *recognize-to-refuse*, not *never-conceive*.