Strong hallucinations from negation and how to fix them

Nicholas Asher
CNRS, IRIT
118 route de Narbonne
Toulouse, France
asher@irit.fr

Swarnadeep Bhar
IRIT / Université Paul Sabatier
118 route de Narbonne
Toulouse, France
swarnadeep.bhar@irit.fr

Abstract

Despite great performance on many tasks, language models (LMs) still struggle with reasoning, sometimes providing responses that cannot possibly be true because they stem from logical incoherence. We call such responses *strong hallucinations* and prove that they follow from an LM's computation of its internal representations for logical operators and outputs from those representations. Focusing on negation, we provide a novel solution in which negation is treated not as another element of a latent representation, but as *an operation over an LM's latent representations that constrains how they may evolve*. We show that our approach improves model performance in cloze prompting and natural language inference tasks with negation without requiring training on sparse negative data.

1 Introduction

This paper investigates and addresses reasoning problems for language models (LMs) that stem from their representation of the meaning of logical terms. Underlying such problems, which we call *strong hallucinations*, are faulty representations that lead LMs into logical incoherence, and hence into error, regardless of the facts. Concentrating on negation, we prove that how an LM computes meaning representations for negation and how that affects outputs inevitably lead to logical error.

We argue that we can eliminate these errors and strong hallucinations by altering the way an LM treats and hence interprets logical operators. We illustrate our approach with negation. Our idea is simple but radical: whereas for LMs all tokens in a context play the same role in determining an attention based probability distribution over next tokens or "continuations" of the context, we claim logical operators should operate differently from non logical expressions. For instance, negation does not contribute just another token to a latent representation via linear algebraic operations; it provides constraints that shift an input probability distribution Π that is determined by the tokens ϕ in its scope to a distribution Π', with $\Pi'(\neg\phi) = 1 - \Pi(\phi)$. These constraints, along with others for sequences or conjunction, allow us to calculate algorithmically and recursively transitions from input distributions over only positive content to appropriate distributions for a continuation containing logical operators. The output distribution not only faithfully reflects the logical meaning of negation but also other first order logical operators.

While hallucinations are often observed for text generation, these settings are difficult to evaluate. Thus, we evaluate our algorithmic or hybrid approach, we call it Λ for negation on three easily evaluable tasks: yes|no question answering, masked knowledge retrieval (MKR) and natural language inference (NLI) with negation, using both encoder LMs and LLMs. Nevertheless, we will illustrate Λ on text generation in NLI.

Λ yields good performance on MKR and NLI; we increase accuracy by 13% on one NLI dataset and by 91% on another. Our datasets feature negation and are modified, linguistically validated versions of (Dagan et al., 2005; Hossain et al., 2020). Λ requires an LM to train and infer only on positive data, which avoids the problem of finding sparse negative data.

Section 2 provides background on hallucinations, distributions and linguistic meaning. Section 3 proves strong hallucinations follow from basic assumptions about an LM's treatment of negation. Section 4 details our positive proposal, and Section 5 provides experiments with our approach on Q&A, MKR and NLI. Section 6 discusses related work.

2 Background

LMs. We assume LMs trained on transformer

Findings of the Association for Computational Linguistics: ACL 2024, pages 12670–12687
August 11-16, 2024 ©2024 Association for Computational Linguistics

architectures (Vaswani et al., 2017) with multiple layers i over very large corpora using masked or causal language modeling. In decoder models, the decoder is trained to predict next token y_i given a context matrix c_i and previously predicted tokens $\{y_1, , y_{i-1}\}$ using a conditional distribution $\mu(y_i/y_1, ..., y_{i-1}, c_i)$. In encoder models, c_i is typically the input to an MLP layer to make predictions. c_i is the LM's internal representation at i of all the input tokens in its context and is a concatenation of multiple hidden states z_j^h, one for each attention head such that:

$$z_j^h = \Sigma_{k=1}^{K} \alpha_{j,k}^h W_V^h x_k \tag{1}$$

$\alpha_{j,k}^h$ are the attention weights providing the importance of token (representation) x_k to x_j and W_V^h is the value weight matrix of head h. All tokens play a similar role given Equation 1.[1]

This representation enables LMs to learn a function $\sigma : (C \times X) \rightarrow [0, 1]$ that assigns probabilities for string $x \in X \subset V^m$ given context $C \subseteq V^n$, where V^n is the set strings of length n with vocabulary V. For encoder and decoder models and tasks we consider here, m = 1 typically. So we will abstract away from architecture and learning details for our models (the tasks we examine all fit in this category); they learn a similar function from contexts to tokens of different kinds, though different loss functions may be used. Generative LMs use conditionalization as in Equation 2 to assign probabilities to novel strings $s = (w_1, w_2, ..., w_{m+1})$ given context C:

$$\mu(s|C) = \mu(w_1|C) \times \mu(w_2\,|C, w_1) \times \tag{2}$$

$$... \times \mu(w_{m+1}|C, w_1, ..., w_m)$$

Generative LMs and encoders use this distribution over strings to select ouputs or continuations given a context C. Various methods $\mathcal{M}$ (maxarg, greedy, beam search, or sampling) over the distribution $\mu(.|C)$ have been proposed to find the maximally plausible continuation $\phi \in X^n$ of C. An LM's affirmations, which we use in defining hallucinations below, are directly linked to this distribution.

Definition 1. $\hat{f}$ affirms ϕ given C, if $\phi = \mathcal{M}\{x \in X : \mu(x|C)\}$

[1]For motivation, see (Baroni and Zamparelli, 2010); these representations contain rich linguistic information (Devlin et al., 2019; Liu et al., 2019; Tenney et al., 2019; Talmor et al., 2020; Zhu and Rudzicz, 2020).

Strong Hallucinations. In this paper, we examine a special kind of hallucination we call *strong hallucinations*. To clarify strong hallucinations, we first provide a gloss for *hallucinations*. Hallucinations involve an attitude like belief or affirmation towards a proposition or sentential content p. An agent A's degree of belief in p, $\mu(p)$, which is dependent on A's evidence E, should determine their sincere assertion of p. This holds for LMs too; they use probability distributions to determine their "affirmations" (output). Now A's degree of belief in p may or may not correlate with a "reasonable" degree of belief in p, $\pi(p)$—a degree of belief that one expects a rational person to have. On one type of hallucination, then, an agent A with grounds or context E hallucinates that p if $\mu_A(p|E)$ diverges significantly from $\pi(p|E)$.

A strong hallucination by A involves a reasoning mistake given evidence. For an LM this evidence is a context or prompt. The result of a strong hallucination is that μ leads A to affirm a claim or set of claims that strongly diverge from any possible reasonable degree of belief given the evidence. A special case is when the set of claims together with the evidence is logically unsatisfiable or cannot possibly be true together.

LLMs have strong hallucinations. Consider, for example this session from LLama2-7b on the NLI task we detail in Section 5. The task was is to decide whether a hypothesis H is entailed by a premise P. To help it with this task, we instructed Llama in a prompt to answer two questions: Q1 *can H be true if P is true?* and Q2 *can H be false if P is true?*. We instructed it to answer "entailment"(E) if it answered no to Q2 – and "neutral" (not E) if it answered yes to Q1 and to Q2. Here is an example of the LLM's response (R):

(R) *"[INST]PREMISE: A young girl rides her bike HYPOTHESIS: This little girl is riding her bike Relation:[/INST] Based on the rules and procedure provided, I will ask myself 1) Can the hypothesis be true if the premise is true and 2) Can the hypothesis be false if the premise is true? Since the premise and hypothesis are the same, it is possible for the hypothesis to be true if the premise is true. Therefore, the answer to question1 is yes. Since the premise and hypothesis are the same, it is not possible for the hypothesis to be false if the premise is true. Therefore, the answer to question 2 is no. Based on my answers to the questions, I will predict a neutral relation between the premise and* *hypothesis" (not E).*

We assume that Llama has taken the context as given with the instructions in which we have if yes to Q1 and no to Q2, then it should infer entailment. Llama answers correctly yes to Q1 and no to Q2, but apparently converts its "no" to Q2 into a yes answer or simply affirms not E despite accepting *A, B and if A and B, then infer E*. This is an example of a logical error that infects not only the LLM's reasoning but also, as in this case, its explanations of its behavior.

We now provide a technical definition of strong hallucinations using probability distributions. We idealize rational distributions by considering *objective* distributions that capture the probabilities of events that are coherent with respect to probability axioms. P represents the objective distribution over events or propositions described by strings in X in the actual state, while P_w is an objective distribution for a logically possible way w the world could be. Let $\mu_{\hat{f}}(.|E)$ be the conditional probability distribution of an LM $\hat{f}$ over a set of truth evaluable strings X that express a set of propositions that includes p.

Definition 2. $\hat{f}$ *strongly hallucinates* that p, if $\mu_{\hat{f}}(p|E)$ significantly diverges from $P_w(p|E)$ for any possible objective distribution P_w and state w.[2]

Probability and linguistic meaning For epistemologists an ideal agent's probability distribution κ governing assertions should equal some objective distribution P (Lewis, 1981; Hall, 1994). An ideal distribution κ thus differs in an important but subtle way from LM distributions. κ is *not* a probability distribution over sentences or strings but over propositions, what sentences express. To link with assertions, we use κ to assign values to strings as follows: for any proposition p, if strings ϕ, ψ both express p, then $\kappa(\psi) = \kappa(\phi)$.In addition, the meanings of logical operators like $\neg$ (not), $\vee$ (or), $\wedge$ (and) impose the structure of a Boolean algebra $\mathbb{P}$ on the set of propositions. Hence, any probability function κ respecting $\mathbb{P}$ that assigns values to sentences of a language verifies all the probability axioms (see Proposition 2 in Appendix A.1) and assigns every sentence expressing a logical truth probability 1, every inconsistent string probability 0 and if ϕ is a logical consequence of ψ then $\kappa(\phi|\psi) = 1$. Conversely, each linguistic model for meaning defines such a probability function κ(See Propositions 3, 4 Appendix A.1). Thus, an agent whose assertions are governed by κ, will avoid strong hallucinations.

3 LMs, negation and strong hallucinations

Section 2 showed a tight connection between ideal distributions, meanings and probability axioms. This connection does not hold for an LM's objective function that is a distribution over strings.

Proposition 1. Every LM $\hat{f}$ whose outputs are governed by $\mu_{\hat{f}}$ and Definition 1 must strongly hallucinate if either: (i) $\sum_{x \in V^n \setminus \{\phi, \neg\phi\}} \mu_{\hat{f}}(x) > 0$; (ii) $\mu_{\hat{f}}$ assigns values to strings in V^n that are logical truths or deductively valid reasoning steps.

Proof: (i). Given the assumptions in (i), $\mu_{\hat{f}}(\phi) + \mu_{\hat{f}}(\neg\phi) < 1$. But for every objective distribution P, $P(\phi) + P(\neg\phi) = 1$. So $\mu_{\hat{f}}$ diverges from any possible objective distribution. (ii). Any objective distribution must assign probability 1 to every logically true sentence ℓ of V^*. For each positive clause Q there is such an ℓ of the form *Not (Q and Not Q)*. As a distribution over individual strings, μ cannot assign 1 to all such ℓ. Moreover, suppose given context the model offers a chain of deductive reasoning, $\phi.\ell.\ell'$. But then by the axioms of probability and Equation 2, an objective distribution gives $P(\ell|\phi).P(\phi) = P(\phi)$. So $P(\ell|\phi) = 1$. If $\mu(\ell|\phi) = 1$, however, μ assigns no other continuation of ϕ non 0 probability; e.g., $\mu(\ell'.\ell|\phi) = 0$. This shows that LM distributions cannot respect the property of objective distributions that logically equivalent formulas receive the same probability. In either case, by Definition 2, the LM strongly hallucinates.□ We note that a similar argument to that in the proof of Proposition 1 (ii) shows that there are pairs of logically equivalent expressions to which $\mu_{\hat{f}}$ will assign distinct values.

(Ramsey, 1931; De Finetti, 1937) linked the probability distributions underlying strong hallucinations to irrational behavior on bet. Our results correctly predict that LMs will place pathological bets (See Appendix G for an example). While most NLP practitioners won't care that their LM bets badly, reasoning errors like that in the example given in Section 1 are serious. Suppose an LM must output a reasoning chain $R = (\phi, \ell_1, ..., \ell_r)$ in which ℓ_i in R should follow deductively from previous elements in R.

[2] In the same spirit, we can define standard hallucinations: $\hat{f}$ *hallucinates* that p, if $\mu_{\hat{f}}(p|E)$ is significantly different from $P(p|E)$ and $\mu_{\hat{f}}(E)$ significantly diverges from $P(E)$.

Proposition 2. If an LM $\hat{f}$ outputs a reasoning chain R with length r, $\mu_{\hat{f}}$ will increasingly diverge from any possible objective distribution and $\hat{f}$ will eventually surely hallucinate as r increases.

Consider a reasoning chain R and suppose $\mu_{\hat{f}}(\ell_1|\phi), \mu_{\hat{f}}(\ell_2|\ell_1), \mu_{\hat{f}}(\ell_2|\phi, \ell_1)$ are high but < 1, which given Proposition 1 is the best we can hope for. By Equation 2 $\mu_{\hat{f}}(\ell_1.\ell_2|\phi) < \mu_{\hat{f}}(\ell_2|\phi, \ell)$. R's probability will thus decrease monotonically as R gets longer. Eventually, for some logically valid continuation of $R.\ell_{r+1}$, $\mu_{\hat{f}}(\ell_{r+1}|R) < \mu_{\hat{f}}(\psi|R)$ for ψ independent of R. So $\hat{f}$ eventually will surely hallucinate.□

All LMs that generate responses from a distribution over individual stringsare subject to Proposition 1 and part (ii) applies to logical operators in general: LMs that make reasoning chains will assert invalid ones or will make mistakes in building them, with the probability of logical error increasing as the chain's length increases.

Our results show we need to rethink the underlying LM distribution over strings to solve the problem of strong hallucinations from negation. While LMs may have some grasp of negation in simple contexts (Gubelmann and Handschuh, 2022), this address the underlying problem with an LM's objective function. We turn to fixing that now with a hybrid approach to logical operators.

4 A hybrid treatment of logical operators for LMs

Logical operators like negation structure meaning recursively by performing a distinctive operation over the content of tokens in its scope. To develop our hybrid treatment of logical operators Λ, we adapt this idea and define a logical operator as function transforming distributions over continuations determined by the latent representations of tokens within the scope of those operators into continuations that reflect the semantics of the operator. For example, negation in the formula $\neg A$ transforms distributions over continuations determined by the latent representation of A into continuations that reflect the semantics of negation and $\neg A$.

To develop Λ, we will make use of a dynamic analysis (DS) of logical operators (Kamp, 1981; Kamp and Reyle, 1993; van Eijck and Kamp, 1997). DS models logical operators as transitions between structures A that are pairs (U_A, P_A). As (Li et al., 2021) have noted, such a pair corresponds to a partition of the tokens in an LM—U_A a set of tokens represents objects and P_A a set of tokens representing properties of those tokens. We take these tokens to be the output of the processing of some input linguistic context C to the LM. A second tool we need is the notion of an embedding (Chang and Keisler, 1973). An embedding of one structure A derived from linguistic information C into another B from another information source $\mathcal{V}$ defines satisfaction or truth of C relative to $\mathcal{V}$.[3]

Definition 3. A has an *embedding* $f : U_A \to U_B$ in B (written $A \leq_f B$) iff for each property ascription $\phi \in P_A$ and for all $x_1, ...x_n \in U_A$ if ϕ holds of or is *satisfied* by $x_1, ...x_n$ in A then ϕ is satisfied by $f(x_1), ...f(x_n)$ in B.

(Kamp and Reyle, 1993) uses embeddings to define the meaning of negation. Negation structurally takes scope over a representation A, and converts it into a property of tokens in a larger structure. Embeddings and their extensions interpret this structural feature. Suppose for some structure C, $A \leq_f C$, we say g *extends* f to an embedding of B in C, if: $g \supseteq f$ and $Dom(g) = U_B$ and $B \leq_g C$.

Definition 4. Let A_ϕ represent ϕ. $A_{\neg\phi}$ is satisfied relative to some $\mathfrak{v} \in \mathfrak{V}$ and map f : iff there is no extension g of f over U_{A_ϕ} such that $A_\phi \leq_g \mathfrak{v}$.

4.1 Negation as a constraint on continuations

Pure LMs can't appeal to an external source to characterize negation's effect on their internal representations. Our key and novel idea uses continuations and their semantics (Asher et al., 2017, 2023). Continuations τ of a string σ have representations A_τ, where: $U_{A_\sigma} \subseteq U_{A_\tau}$ and $P_{A_\sigma} \subseteq P_{A_\tau}$ (which we write as $A_\sigma \preceq A_\tau$). Then trivially for the identity map $\iota : U_{A_\sigma} \to U_{A_\sigma}$, $A_\sigma \leq_\iota A_\tau$. Conflating continuations and the structures representing them, a continuation A_2 of a context A_1 has, given an LM's distribution μ, a probability α–written $A_1 \preceq_\alpha A_2$; α is just $\mu(A_2|A_1)$.[4]

A proper treatment of negation determines how content in the scope of the negation affects coherent or admissible distributions of continuations A_2

[3]Our formulation here echoes maps used in multimodal approaches (Li et al., 2019; Lu et al., 2019; Driess et al., 2023; Devillers et al., 2023; VanRullen and Kanai, 2021).

[4]Note that α is not the probability of embedding A_1 in A_2; that probability should always be 1, as ι always exists and provides the embedding.

of a context A_1. For example, if an object (or token representation) o has the property $\neg A$ in A_1, no consistent or coherent continuation of the list of properties for o in A_2 should contain A. More precisely, Negation imposes two constraints on coherent or *admissable* continuations A_2 and distributions μ of an LM relative to a context A_1; the third constraint below forces μ to obey the logical interpretation of conjunction or concatenation in strings; while the introduction of new information in a continuation, which is a way of representing conjunction introduction, is naturally represented in terms of conditionalization as in equation 2, constraint (iii) below ensures that an LM uses conjunction elimination correctly. For structures A, B, let $A + B = (U_A \cup U_B, P_A \cup P_B)$

Definition 5. Let $A_1 \preceq_\gamma A_2$ with μ an LM. A_2 is an *admissible continuation* of A_1 relative to μ only if:
(i) If $\{\neg A_k\} \cup P_{A_1} = P_{A_2}$ with $\mu(A_k|A_1) = \alpha$ then $\gamma = 1 - \alpha$;
(ii) if $P_{A_1} = \{\neg A_k\}$ and $A_3 \prec A_2$ and $P_{A_2} = P_{A_1} \cup P_{A_3}$, then[5] $\gamma = \frac{\mu(A_3)(1-\mu(A_k|A_3))}{1-\mu(A_k)}$;
(iii) if $\gamma = 1$, then for any context C, $\mu(A_2|C) \leq \mu(A_1|C)$;
(iv) for any $A_4 \preceq A_2$ (A_4 possibly empty), $\mu(A_1 + A_4|A_2) = 1$.

Definition 5 constrains the evolution of an LM's latent representation A and distribution to capture the meanings of logical operators like negation in A in terms of their information update potential, in terms of what distributions over continuations they permit as admissible.

To illustrate, suppose a distribution μ with $A_1 \preceq_\gamma A_2$, A_2 represents *it is not the case that* B and $\mu(B|A_1) = \alpha$. Definition5 says that A_2 is an admissible continuation with respect to μ only if $\gamma = 1 - \alpha$. Negation maps $\mu(B|A)$ into $1 - \mu(B|A)$. For any context C, unless $\mu(B|C) = .5$, $\mu(B|C) \neq \mu(\neg B|C)$.

Definition 5 applies recursively to strings of the form $\neg(\phi \wedge \neg\phi)$ to assign them probability 1 and to $\phi \wedge \neg\phi$ to assign them 0, regardless of the probability of ϕ. We can also model conditionals $A \Rightarrow B$, using Definition 5, by translating *if* ϕ *then* ψ as $\neg(A_\phi.\neg A_\psi)$. So for any context C $\mu(A_\phi \Rightarrow A_\psi|C) = 1 - \mu(A_\phi|C) + \mu(A_\psi|A_\phi + C) \times \mu(A_\phi|C)$, which in turn validates modus ponens: $\mu(B|A \Rightarrow B.A.C) = 1$. So, given a context C, as in the strong hallucination example R of Section 2, that contains the conditional *(yes to Q1 and no to Q2)* $\Rightarrow$ *Entailment* as well as a *yes* answer to Q1 and a *no* to Q2, then $\mu(Entailment|C) = 1$. Definition 5 also constrains LM objective functions to conform to the truth conditional meanings of propositional connectives and quantifiers. Like negation, quantifiers operate on embeddings. The universal quantifier has the same meaning as that of the conditional we gave above (Kamp and Reyle, 1993; Groenendijk and Stokhof, 1991), while the meaning of an existential quantifier is captured by the definition of an embedding operating on the elements of U_A that function as variables. For details see Appendix A.5.

[5]For details see Appendix A.4.

5 Experiments with our approach on Q&A, cloze prompts, NLI and negation

We illustrate our hybrid approach Λ of Section 4 on three applications that feature negation: simple Q&A, masked knowledge retrieval (MKR) and natural language inference (NLI), each with a different output and input. We use both encoder and decoder LMs and LLMs.

These applications require us to apply Λ's constraints logical operators, in particular negation, at various points. Suppose we must predict strings in $\mathcal{S}$. $\mathcal{S}$ will have strings or substrings $\mathcal{S}'$ free of logical operators like negation or conditionals (e.g. simple clauses). Λ leverages an LM's distribution μ_1 over $\mathcal{S}'|\mathcal{I}'$ where $\mathcal{I}'$ is also free operator free and then transforms μ_1 using the constraints in Definition 5 into a distribution over $\mathcal{S}|\mathcal{I}$, which contain operators.

While Λ generalizes to all LLM tasks, our applications only require predicting very short strings. While tasks with long string output are beyond this paper's scope because of the complex evaluations required(Li et al., 2023), we will show how Λ can improve the longer reasoning chains of the LLM when prompted on NLI.

Q&A tasks. We examined BERT large and basic encoder models on a question answering task about facts in a synthetic dataset SYN we made. We used only Y|N questions with positive and negative contexts like: *there was a* $< col >$ *car* (positive context); *there was no* $< col >$ *car* (negative context). "$< col >$" is a placeholder for a color term. Given such contexts, we asked, Was there a $< col >$ car? (See Appendix B for details).

Given an input $< cls >$ `QUESTION` $< sep >$ `CONTEXT` $< sep >$ and a BERT model, we plotted cosine similarity (cossim) histograms of CLS representations from a positive context and from its corresponding negative context, before and after fine tuning using the training regime of (Chaturvedi et al., 2024) on the Coqa dataset (Reddy et al., 2019). We used BERT, as its pretraining gives a meaningful CLS representation (Devlin et al., 2019). On pretrained BERT, cossim values ranged between .986 and 1, showing that, as one might expect, pretrained BERT's representations do not give appropriate content to negation.

After Q&A fine tuning, however, cossim values for the CLS representations of positive and negative contexts were much lower, indicating that both BERT models (especially BERT-large with values between 0,34 and 0,38) learned to differentiate positive and negative content in CLS tokens. (Histogram plots are in Appendix C.)

We then tested fine-tuned BERT and Roberta on our Q&A task; the large models did perfectly though the small ones mostly gave only a "no" answer. The models interpret our task as a binary classification, in which the model considers only two mutually exclusive and exhaustive possible continuations, P (for "yes" to P?) and $\neg P$ (for "no" P?. This strategy approximates the truth functional meaning of negation in a simple setting and might account for their success with negative contexts.

Finally, we used Λ and fine-tuned BERT and RoBERTa models only on positive contexts in SYN. We accessed the logits, which after softmax give us a probability distribution over the Y|N classes. Using the probabilities over the positive contexts $(Y|?Q, C)$, we applied Definition 5 to get probabilities for outputs with negative responses or contexts–e.g., $(N|?Q, C)$ or $(Y|?Q, \neg C)$. Λ matched the esults of the large models on our task and yielded the best consistent results for the small models.

Masked Knowledge Retrieval. In a second experiment, we explored a Masked Knowledge Retrieval (MKR), in which the model must compute a sentence completion and lexical item for *mask* in context C– where the mask is either in a positive C or negative $\neg C$. Regardless of the facts, a prediction for $\neg C(< mask >)$ should not match the prediction for $C(< mask >)$; answers preferred in C should be dispreferred in $\neg C$.

We used (Kassner and Schütze, 2020)'s (KS) and (Jang et al., 2022a)'s (JS) "negated" versions of the LAMA dataset (Petroni et al., 2019). We converted the negative contexts for the MKR task provided in KS and JS into a positive context C, which we need for our approach. KS provided 51 prompt examples, while JS yielded 2926 examples.

Pretrained RoBERTa-large and BERT-large had significant numbers of examples with an exact match (EM) for both positive and negative contexts in KS and so showed an inconsistent treatment of negation and strong hallucination. For instance, given the positive (negative) contexts, *A teacher is (not) most likely teaching at a [mask]*, they returned the completions, *A teacher is most likely teaching at a [school]* and *A teacher is not most likely teaching at a [school]* (See Appendix D for more examples). Finetuning RoBERTa and BERT on Q&A as above improved model performance reducing EM but not eliminating them (see Table 1). We concluded that fine-tuning on Q&A does not give a full understanding of negation.

With cloze prompts, an LM originally outputs only the token with the highest score. But an important feature of negation is that it presupposes a set of relevant alternatives (Rooth, 1992). *The capital of France is not Marseille* conveys the information that while the capital of France is not Marseille it is some other city or place where people live and work.

To capture the idea of relevant alternatives with Λ, we modified the LM's output to get its top 5 candidate completions in its distribution. We then applied Definition 5 by computing the probabilities p_i over the top 5 completions σ_i $1 \leq i \leq 5$ of the positive prompt $C(< mask >)$; for the negative contexts $\neg C(< mask >)$, we assigned $1 - p_i$ to σ_i, reversing the LM's ranking of positive completions.

Λ produced 0 EM on both KS and JS. All completions were meaningful on KS and only relatively few were ungrammatical on JS. Table 1 gives the performance on MKR of BERT and RoBERTA with pre-training only, fine tuning on Q&A and Λ.

Model	Dataset	Pre-t	FT-CoQA	Λ
RoBERTa-L	KS	32/51	10/51	0 / 0
	JS	1038/2926	743/2926	0/6
BERT-L	KS	30/51	17/51	0/ 0
	JS	970/2926	814/2926	0/162

Table 1: MKR Accuracy for Roberta-large and BERT-large with pre-t(raining only)/fine-tuned with CoQA (FT-CoQA) and Λ. For Pre-tr and FT, we give #EM / # examples. For Λ we give #EM /# non meaningful completions.

NLI. In a third application, we looked at natural language inference (NLI) with RTE (Dagan et al., 2005) and SNLI (Bowman et al., 2015), two datasets modified for negation by (Hossain et al., 2020). Each contains a context C and a hypothesis h, which are labeled either with entailment (which we note as (C, h):E) or non-entailment $\neg E$ (RTE) or entailment (E), contradiction (Cn) or neutral (N) (SNLI). (Hossain et al., 2020) negated manually either C or h or both in portions of RTE and SNLI. Hossain's datasets, ¬RTE and ¬SNLI, contain no examples of positive cases (C, h).We reconstructed positive examples (41% of ¬RTE) and for ¬SNLI; and to study negation in more detail we completed 117 examples of (Hossain et al., 2020)'s ¬SNLI with four inferential patterns: (C, h), $(C, \neg h)$, $(\neg C, h)$ and $(\neg C, \neg h)$.

In addition, in ¬SNLI, negations of (C, h) pairs were often wrongly labeled because of non full scope negation. For instance, there were examples where the original h was of the form *a man is smoking* and a context C such that $(C, h) : E$ was correct; but where with the negative hypothesis $\neg h$ was *a man was not smoking* we had $C, \neg h$: Cn, when C and $\neg h$ were plainly consistent and we should have had $C, \neg h : N$. We made either changed the label or made h and $\neg h$ true contradictories, using a definite description *the NP* in h to refer to *a NP* from C or replacing *a NP* in h with *no NP* (for examples see Appendix E).

In addition our data often featured less than full sentential scopes for negation on C, as well as presuppositional elements (definite noun phrases, adverbial clauses), which scope out of negation. To deal with this, we divided C into two components P and C', with negation having scope only over C'. We then added annotated inferential patterns for both material under the scope of negation and material outside of it. For instance, if we had the context *a man with blue shoes was not sleeping* and a hypothesis h, we annotated the inference to h from *a man with blue shoes* (P, h), and from h to *a man was not sleeping* (h, C'). Our datasets containing 2306 annotated inferences are at `https://github.com/swarnadeep8597/Negation.git`.

In NLI asks, the LM must learn to label context and hypothesis with relations whose definitions involve negation. This imposes a logical structure on the labels given their intuitive meaning; e.g., entailment (E) between C and h means that there are no situations where C and $\neg h$ hold. From the meaning of entailment and Definition 5, it follows that $\mu(E|C, \neg h) = 1 - \mu(E|C, h)$. Thus, if the model predicts (C, h):E, it should predict $(C, \neg h)$: $\neg E$. To infer labels for $(\neg C, \neg h)$ from positive data, if (h, C):$\neg E/E$, then $(\neg C, \neg h)$: $\neg E/E$. Similarly for $(\neg C, h)$ contexts: if $\neg C, \neg h$:E, then $(\neg C, h)$:$\neg E$. This illustrates the interactions between negation in the data and in the task definition. Our rules are provably correct but incomplete for ¬RTE; using only valid rules, we cannot infer $(C, \neg h)$:E from just positive data in the 2 label E and $\neg E$ task. So our method perforce missed those cases (see results for Λ on $C, \neg h$ in Table 2 for ¬RTE).

To account for less than wide scope negations, we developed an algorithm inspired by (Karttunen and Peters, 1979) that determines values for all configurations based on entailment predictions for (C, h), (P, h) and (h, C'). In Algorithms 1 and 2 are the scoped algorithm for $(\neg C, h)$ and $(\neg C, \neg h)$ in a 2 label NLI problem (the ¬RTE dataset) and scoped C. As we did not scope h, only those cases are relevant.

Algorithm 1 Algorithm for $(\neg C, h)$ and a 2 label NLI problem

1: **procedure** ALGORITHM$(\neg C, H)$
2: (C, h) :← defined in paper
3: (P, h) :← P defined in paper
4: **if** $(h, C) : \neg E$ **then**
5: **if** $(C, h) : E$ and $(P, h) : E$ **then**
6: $(\neg C, h) : E$
7: **else**
8: $(\neg C, h) : \neg E$
9: **if** $(h, C) : E$ **then** $(\neg C, h)$: $\neg E$

In our ¬SNLI, the three labels E, Cn and N also make reference to logical relations involving negation, which Λ must reflect in its algorithm. For instance, exploiting these relations for $C, \neg h$ contexts gives: (C, h):$E/Cn/N$ iff $(C, \neg h)$:$Cn/E/N$. Provably correct and complete rules for scoped negation in the 3 label NLI task are in appendix F, capturing the cases we missed in RTE.

Table 2 contains scores for the different configurations of C, h and negation. We first looked at finding a theoretical maximum accuracy for Λ using gold labels for for (C, h), (P, h) and (h, C') labels on both our ¬RTE and ¬SNLI. For our ¬RTE, Λ yielded an accuracy of 94% on all

Algorithm 2 Algorithm for $(\neg C, \neg h)$ and a 2 label NLI problem

procedure ALGORITHM$(\neg C, \neg H)$
    (C, h) $:\leftarrow$ C,h defined in paper
    (P, h) $:\leftarrow$ P defined in paper
    (h, C') $:\leftarrow$ C' defined in paper
    if $(h, C) : E$ **then**
        if $(P, h) : E$ **then**
            $(\neg C, h) : E$
        else
            $(\neg C, h) : \neg E$
    if $(h, C) : \neg E$ **then**
        if $(h, C') : E$ **then**
            $(\neg C, h) : E$

C,h configurations, while Λ basic, an algorithm assuming wide scope negation over C, was 11% less accurate, showing the importance of scoping (Kletz et al., 2023). On our ¬SNLI, Λ yielded a 96% accuracy overall.

Next, we prompted Llama2 7B on our NLI tasks to compute labels for positive (C, h), (P, h) and (h, C') data, to be used with Λ. We experimented with different formats; the model did best with prompts that involved using answers to two Y|N questions to infer a label in a chain of thought style (Wei et al., 2023), an example of which is in (R) (Section 2). Longer prompts for (C, h), shorter ones for (P, h) and (h, C') worked best are given in https://github.com/swarnadeep8597/Negation.git. The prompts in the style of (R) provided extended reasoning chains with conditionals; and Definition 5 enabled us to exploit Λ also for the positive cases. Llama produced about 12% incorrect reasoning chains that Λ could correct. However, because the valid inferences led to faulty labels, the scores did not go up. So we took Llama's labels for the (C, h), (P, h) and (h, C') cases.

Table 2 gives results for (LΛ) and and Llama alone (L). On our ¬RTE, for the positive datasets, L got 73% correct for (C, h), 76% correct on (P, h) and 83% on (h, C) labels. We used these predictions so that L, Λ achieved an average overall accuracy of 80% on the various $((\neg)C, (\neg)h)$ configurations. Llama alone (L) had an average accuracy of 71% on all $((\neg)C, (\neg)h)$ configurations–with 76% on $(C, \neg h)$, 63% for $(\neg C, h)$ and 70% for $(\neg C, \neg h)$. $L\Lambda$ significantly outperformed Llama by itself on this task.

L did less well on our ¬SNLI as seen in Table 2. After several attempts we found prompts on which$L$ predicted 78% correct on (C, h), 69% on (P, h) and 70% on (h, C'). Using these scores $L\Lambda$ gave an overall accuracy of 72%. L achieved a decent accuracy on $(C, \neg h)$ (accuracy 71%); but floundered on the $(\neg C, h)$, and $(\neg C, \neg h)$ cases. $L\Lambda$ far surpassed L on the NLI task with negation.

Data	Env.	$C, \neg h$	$\neg C, h$	$\neg C, \neg h$	Full
¬RTE	Λ basic	.89	.76	.91	.85
	Λ	.89	.96	.98	.94
	L Λ(.73)	.89	.76	.76	.8
	L (.73)	.76	.63	.70	.71
¬SNLI	Λ	.96	.97	.95	.96
	L Λ (.78)	.77	.67	.67	.72
	L (.78)	.71	.12	.11	.43

Table 2: Accuracy on NLI tasks for ¬RTE and ¬SNLI datasets. $\Lambda basic$ accuracies for basic algorithm assuming sentence wide scope. Λ: accuracies for the full algorithm with scoping on the (¬)C,(¬)h configurations, given gold labeled (C, h), (h, C'), $P, h)$. L: Llama predictions with best prompts. LΛ: predictions using Λ given $L\Lambda$ predictions for (C, h), (h, C'), $P, h)$.

6 Related work

Definition 2 of strong hallucinations shows they involve necessarily unfaithful content like what (Ji et al., 2023) call *instrinsic hallucination* in which the unfaithful content contradicts its source. Strong hallucinations contradict every possible source (Moramarco et al., 2022; van Deemter, 2024). While researchers have proposed various causes for hallucinations (Filippova, 2020; Parikh et al., 2020; Longpre et al., 2021), we are, as far as we know, the first to define and to analyze strong hallucinations and to derive strong hallucinations from the distribution governing an LM's output.

Various proposals to avoid hallucinations have also surfaced (Nakano et al., 2021; Asher and Hunter, 2022; Merrill et al., 2022; Gubelmann and Handschuh, 2022). The most obvious proposal, building bigger LLMs and larger training corpora, has drawn criticism (Filippova, 2020; Huang et al., 2021; Li et al., 2023; Goyal and Durrett, 2020; Sellam et al., 2020; Li et al., 2023). Our negative results buttress this criticism. Our method for correcting strong hallucinations from negation is formally correct; it's not a form of training but acctually shifts LM output in its last layer. The embeddings we use to define Λ could serve to check factual hallucinations as they refine and extend lexical matching and relation extraction techniques (Dhingra et al., 2019; Cao et al., 2018; Huang et al., 2020).

Prior research (Kassner and Schütze, 2020; Hossain et al., 2020; Hosseini et al., 2021; Jang et al., 2022a) has reported mediocre results for LM performance with negation on MKR and NLI. Proposition 1 provides a formal foundation for these observations, and our tests on pretrained BERT models confirms that they doesn't adequately represent negation. (Gubelmann and Handschuh, 2022) adopt a pragmatic approach that improves LM performance on MKR; our novel, semantic approach gives completely logically coherent MKR performance. Finally, (Truong et al., 2023) investigated GPT style LLMs and found NLMs unable to reason with negation. Our prompting of Llama2 7B LLM our NLI datasets concurs with their observations.

(Jang et al., 2022b) improve LM performance on MKR tasks with negation by training the model on an intermediate task, in which they train an encoder model to predict whether given an input word and sentence the sentence describes it. They then use the model on MKR tasks. We believe their binary classification task can help understand negation, as our Q&A experiments indicate.

Work on neurosymbolic models (Poole, 2011; De Raedt et al., 2020; Olausson et al., 2023) is also relevant to our approach. Λ leverages the large store of lexical and conceptual knowlwedge in LLMs while constraining the LLM's objective function to comply with the meanings of logical operators. Other neurosymbolic approaches try to incorporate logical operators into the architecture (using logical gates) without changing the objective function (Riegel et al., 2020). Our negative results show that such neuro-symbolic approaches are problematic. More promising post hoc methods use an LLM to translate natural language into some formalism that symbolic methods like theorem provers can manipulate (Olausson et al., 2023) or use LLMs to generate additional symbolic rules (Kalyanpur et al., 2022). However, using LLMs only as translators means we can't access their deep lexical knowledge.

Our paper's negative results assume a learned distribution over strings from corpora. This is compatible with results about what algorithms an LLM can compute when this distribution is replaced with something else (Pérez et al., 2021; Chiang et al., 2023).

7 Conclusions

We have investigated strong hallucinations in LMs, originating from faulty meaning representations of logical operators. Strong hallucinations account for logical errors and the failure of generative LLMs as in the example of Section 2 to explain their own predictions even if those predictions are correct. We proved that under minimal assumptions strong hallucinations must result from an LM's distributions defined over strings. The received view of LM distributions is thus not tenable.

Focusing on negation, we proposed a new treatment Λ of logical operators, on which the operator introduces an operation over latent representations, imposing constraints on how those distributions over those representations evolve in coherent continuations. Our approach reinterprets an LM's distribution as an assignment of degrees of truth to propositions or sets of strings. The constraints will interact with algorithms for tasks that exploit logical structure.

We illustrated Λ on Q&A, sentence completion and NLI tasks, with simple and precise evaluation metrics, and showed that it can increase LM performance. Λ only requires LM training on positive data, which is an advantage, since negative data is sparse in real corpora. For NLI, we tested Λ on two new datasets for NLI with negation that modify those in (Hossain et al., 2020); we showed Λ substantially improved LLM performance on NLI label prediction.

We did not fine tune our models on NLI or train them to predict logical operator scope (Kletz et al., 2023). We need to do both for end to end good performance. Llama2-7b was highly unstable under prompting and failed to grasp the NLI problem well. We hope fine tuning will give better results. We also plan to use the newer and more stable Llama 3 models on our NLI task.

Another task for future work is to deal with the fact that with Λ a distribution may generate a high probability for strings that may not be pragmatically relevant or appropriate. We will explore how to supplement and refine the distribution with pragmatic constraints that are learned with RLHF methods (Mishra et al., 2022; Rafailov et al., 2024). If RLHF has to be done anyway to get good performance from LLMs, perhaps this is not such an onerous step. We also plan to use new annotated data sets (Li et al., 2023) to apply Λ to multiple reasoning steps in LMs.

Acknowledgments

We thank Julie Hunter and Akshay Chaturvedi of the COCOBOT and COCOPIL teams for helpful comments on prior drafts of this paper. For financial support, we thank the ANR project COCOBOTS (ANR-21-FAI2-0005), theANR/DGA project DISCUTER (ANR- 21-ASIA-0005), the COCOPIL "Graine" project of the Région Occitanie of France and the National Interdisciplinary Artificial Intelligence Institute ANITI (Artificial and Natural Intelligence Toulouse In- stitute), funded by the French 'Investing for the Future–PIA3' program under the Grant agreement ANR-19-PI3A-000. This work was performed using HPC resources from CALMIP (Grant 2016-P23060).

8 Limitations

While we have provided empirical support for our approach Λ using both LMs and LLMs on problems with simple and precise evaluation metrics, we haven't deeply investigated negation in full text generation. We had trouble getting Llama even to produce reasoning chains and explanations of its behavior in NLI. Moreover, while it produced about a 10% error rate in those reasoning chains, when we corrected the reasoning the conclusions were often faulty because of Llama's faulty inferences on positive NLI contexts.

Λ doesn't get perfect scores on our NLI data because it doesn't capture the complex interactions between scopes of multiple logical operators. The scope of operators is crucial as can be seen from our Table 2. We did not train our models to do scoping. in future work we will do this following (Kletz et al., 2023). Preliminary experiments show LLama relatively capable of providing logical forms with plausible scopes for logical operators.

Finally, we need to address the failings of LLMs in dealing with just positive data. We need to fine tune a Llama model rather than simply prompting it, as prompting is not a very robust but rather unstable process on these LLMs we have found. To do fine tuning, we will need more and better data. Our examination of the data on NLI shows that we need a much more comprehensive and varied dataset. We will continue to enlarge the revised datasets we have provided here in supplementary material.

References

Nicholas Asher, Swarnadeep Bhar, Akshay Chaturvedi, Julie Hunter, and Soumya Paul. 2023. Limits for learning with large language models. In *12th Joint Conference on Lexical and Computational Semantics (*Sem)*. Association for Computational Linguistics.

Nicholas Asher and Julie Hunter. 2022. When learning becomes impossible. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 107–116.

Nicholas Asher, Soumya Paul, and Antoine Venant. 2017. Message exchange games in strategic conversations. *Journal of Philosophical Logic*, 46.4:355–404.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1183–1193.

Patrick Blackburn, Maarten De Rijke, and Yde Venema. 2001. *Modal logic, Cambridge Tracts in Theoretical Computer Science No.53*. Cambridge University Press.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Chen Chung Chang and H Jerome Keisler. 1973. *Model theory*. North Holland, Elsevier.

Akshay Chaturvedi, Swarnadeep Bhar, Soumadeep Saha, Utpal Garain, and Nicholas Asher. 2024. Analyzing Semantic Faithfulness of Language Models via Input Intervention on Question Answering. *Computational Linguistics*, pages 1–37.

David Chiang, Peter Cholak, and Anand Pillay. 2023. Tighter bounds on the expressivity of transformer encoders. In *International Conference on Machine Learning*, pages 5544–5562. PMLR.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Bruno De Finetti. 1937. La prévision : ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, 7:1–68.

Luc De Raedt, Sebastijan Dumančić, Robin Manhaeve, and Giuseppe Marra. 2020. From statistical relational to neuro-symbolic artificial intelligence. *arXiv preprint arXiv:2003.08316*.

Benjamin Devillers, Léopold Maytié, and Rufin VanRullen. 2023. Semi-supervised multimodal representation learning through a global workspace. *arXiv preprint arXiv:2306.15711*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*.

Haim Gaifman. 1964. Concerning measures on boolean algebras. *Pacific Journal of Mathematics*, (14):61–73.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. *arXiv preprint arXiv:2010.05478*.

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy*, 14(1):39–100.

Reto Gubelmann and Siegfried Handschuh. 2022. Context matters: A pragmatic study of plms' negation understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4602–4621.

Ned Hall. 1994. Correcting the guide to objective chance. *Mind*, 103(412):505–517.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. *arXiv preprint arXiv:2005.01159*.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Mj Jang, Frank Mtumbuka, and Thomas Lukasiewicz. 2022a. Beyond distributional hypothesis: Let language models learn meaning-text correspondence. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2030–2042.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022b. Becel: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Aditya Kalyanpur, Tom Breloff, and David A Ferrucci. 2022. Braid: Weaving symbolic and neural knowledge into coherent logical explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10867–10874.

H. Kamp and U. Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers.

Hans Kamp. 1981. A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, and M. Stokhof, editors, *Formal Methods in the Study of Language*, pages 277–322. Mathematisch Centrum, Amsterdam.

Lauri Karttunen and Stanley Peters. 1979. Conventional lmplicature. In *Presupposition*, pages 1–56. Brill.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818.

David Kletz, Marie Candito, and Pascal Amsili. 2023. Probing structural constraints of negation in pretrained language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 541–554.

David Lewis. 1981. A subjectivist's guide to objective chance. In *IFS: Conditionals, Belief, Decision, Chance and Time*, pages 267–297. Springer.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

William Merrill, Alex Warstadt, and Tal Linzen. 2022. Entailment semantics can be extracted from an ideal language model. *arXiv preprint arXiv:2209.12407*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. pages 3470–3487.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Aleksandar Savkov, and Anja Belz. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *ACL 2022: 60th Annual Meeting of the Association for Computational Linguistics*, pages 5739–5754. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Theo X Olausson, Alex Gu, Ben Lipkin, Cedegao E Zhang, Armando Solar-Lezama, Joshua B Tenenbaum, and Roger P Levy. 2023. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Jorge Pérez, Pablo Barceló, and Javier Marinkovic. 2021. Attention is turing-complete. *Journal of Machine Learning Research*, 22(75):1–35.

F. Petroni, T. Rocktäschel, A. H. Miller, P. Lewis, A. Bakhtin, Y. Wu, and S. Riedel. 2019. Language models as knowledge bases? In *In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019*.

David Poole. 2011. Logic, probability and computation: Foundations and issues of statistical relational ai. In *International Conference on Logic Programming and Nonmonotonic Reasoning*, pages 1–9. Springer.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Frank Plumpton Ramsey. 1931. *The foundations of mathematics and other logical essays*. K. Paul, Trench, Trubner & Company, Limited.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Ryan Riegel, Alexander Gray, Francois Luus, Naweed Khan, Ndivhuwo Makondo, Ismail Yunus Akhalwaya, Haifeng Qian, Ronald Fagin, Francisco Barahona, Udit Sharma, et al. 2020. Logical neural networks. *arXiv preprint arXiv:2006.13155*.

M. Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Alfred Tarski. 1944. The semantic conception of truth: and the foundations of semantics. *Philosophy and phenomenological research*, 4(3):341–376.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan

Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: An analysis of language models on negation benchmarks. *arXiv preprint arXiv:2306.08189*.

Kees van Deemter. 2024. The pitfalls of defining hallucination. *Computational Linguistics*, pages 1–10.

J. van Eijck and H. Kamp. 1997. Representing discourse in context. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Linguistics*, pages 179–237. Elsevier.

Rufin VanRullen and Ryota Kanai. 2021. Deep learning and the global workspace theory. *Trends in Neurosciences*, 44(9):692–704.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Zining Zhu and Frank Rudzicz. 2020. An information theoretic view on selecting linguistic probes. *arXiv preprint arXiv:2009.07364*.

Appendix

A.1 Language, probabilities and propositions 4 and 5

To show the connection between probability and logical truth, we will use the notion of a modal model (Blackburn et al., 2001). Modal models consist of a set of points of evaluation and an interpretation function assigning appropriate meanings to tokens. Let $\mathfrak{A}$ be a modal model (Blackburn et al., 2001) with a set of possible worlds or points of evaluation $W_{\mathfrak{A}}$. For $w \in W_{\mathfrak{A}}$, we write $\mathfrak{A}, w \models \phi$ for the fact that ϕ is true in $\mathfrak{A}$ at w, and $\|\phi\|_{\mathfrak{A}} = \{w \in W_{\mathfrak{A}} : \mathfrak{A}, w \models \phi\}$.

Definition 6. Let $\phi, \psi \in Sent(\mathcal{L})$. ϕ is a semantic or logical consequence of ψ ($\psi \models \phi$) if for all models $\mathfrak{A}$ and worlds $w \in W_{\mathfrak{A}}$ such that $\mathfrak{A}, w \models \psi$, then $\mathfrak{A}, w \models \phi$ (Chang and Keisler, 1973).

Proposition 3. Let P be a probability distribution that respects the σ algebra $\mathbb{P}$. Then P determines a probability distribution π over $Sent(\mathcal{L})$ such that the following axioms A hold: $\pi(\neg\phi) = 1 - \pi(\phi)$; If $\models \phi$, then $\pi(\phi) = 1$; if $\phi \models \psi$, then $\pi(\phi) \leq \pi(\psi)$; if $\models \neg(\phi \wedge \psi)$, then $\pi(\phi \vee \psi) = \pi(\phi)+\pi(\psi)$; $\|\phi\wedge\psi\| = \|\phi\| \cap \|\psi\|$, and $\pi(\forall x\phi) = lim_{n\to\infty} \bigwedge_{a_1,\ldots a_n \in C_{\mathcal{L}}} \pi(\phi(\frac{x}{a}))$.

A straightforward consequence Proposition 3 is that π must assign the same probability to sentences ϕ, ψ if $\models \phi \leftrightarrow \psi$ (are logically equivalent).

Finally, we can relate probability distributions π to models of linguistic meaning. Let $\mathcal{L}$ be a first order language.

Proposition 4. (i) Every probability function $\pi : Sent(\mathcal{L}) \to [0, 1]$ that respects $\mathbb{P}$ defines a modal model $\mathfrak{A}$ of $\mathcal{L}$ with worlds W where: if ϕ is a logical truth, $\|\phi\|_{\mathfrak{A}} = W$; if ϕ is inconsistent, then $\|\phi\|_{\mathfrak{A}} = \varnothing$; If $\models \phi \leftrightarrow \psi$, then $\|\phi\|_{\mathfrak{A}} = \|\psi\|_{\mathfrak{A}}$ and all identities of $\mathbb{P}$ hold in $\mathcal{A}$; finally if $\phi \models \psi$, $\pi(\psi|\phi) = 1$. (ii) Every model of $\mathcal{L}$ defines a probability function $\pi : Sent(\mathcal{L}) \to [0, 1]$ verifying the axioms A of Proposition 3, and (iii) if $\pi : Sent(\mathcal{L}) \to [0, 1]$ does not verify the axioms A, it does not define a model of $\mathcal{L}$, or preserve semantic consequence.

Proof: To prove (i), we consider a set of worlds W where π provides a uniform distribution over $w \in W$. If $\pi(\phi) = \alpha$, we build a model $\mathfrak{A}$ such that where $\|\phi\| = \{w \in W : \mathfrak{A}, w \models \phi\}$, $\sum_{w\in\|\phi\|} \pi(w) = \alpha$. Now suppose ϕ is a logical truth. Then $\pi(\phi) = 1$. So $\sum_{w\in\|\phi\|} \pi(w) = 1$. Given that π is a probability measure over W, $\|\phi\| = W$. Similarly if ϕ is inconsistent, then $\pi(\phi) = 0$. And so $\|\phi\| = \varnothing$. Given the axioms in A, if $\models \phi \to \psi$, then $\pi(\phi) \leq \pi(\psi)$; we then set $\|\phi\| \subseteq \|\psi\|$ in $\mathfrak{A}$. So if $\models \phi \leftrightarrow \psi$, $\|\phi\| = \|\psi\|$. This means that all the identities of $\mathbb{P}$ and all logical equivalences hold in $\mathfrak{A}$. Finally, if $\phi \models \psi$, $\|\phi \wedge \psi\| = \|\psi\|$ given our definition of $\mathfrak{A}$. Then $\pi(\phi|\psi) = \frac{\pi(\|\phi\| \cap \|\psi\|)}{\pi(\|\psi\|)} = \frac{\pi(\|\psi\|)}{\pi(\|\psi\|)} = 1$.
To prove (ii), assume a modal model $\mathfrak{A}$ and define a set of measures Π on $W_{\mathfrak{A}}$ such that $\pi \in \Pi$, $\pi(W_{\mathfrak{A}}) = 1$; $\pi(\varnothing) = 0$; and for $p, q \subseteq W_{\mathfrak{A}}$, if $p \subset q$ then $\pi(p) \leq \pi(q)$; and if $p \cup q = 0$, then $\pi(p \cup q) = \pi(p) \cup \pi(q)$. Such measures exist (Gaifman, 1964) and in virtue of Proposition **??** obey axioms A.
To prove (iii), if π does not obey axioms A, then one of the clauses of satisfaction for L formulas (Tarski, 1944) will fail; since a model of $\mathcal{L}$ must by definition satisfy those clauses, π cannot define a model. □

Proposition 5. Every ideal distribution defines a modal model of $\mathcal{L}$.

Proof: By hypothesis the objective probability distribution P respects the structure of $\mathbb{P}$ and thus P obeys the axioms in A. By Proposition 4 P defines a model $\mathfrak{A}_P$ of $\mathcal{L}$. By definition of κ, $\kappa(\phi|T) = P(\phi)$, so $\kappa(.|T)$ defines a model $\mathfrak{A}_\kappa$ of $\mathcal{L}$ that is elementarily equivalent to $\mathfrak{A}_P$ (Chang and Keisler, 1973). But this means that κ must respect the structure of $\mathbb{P}$ and has the requisite properties via Proposition 4 to define a model

A.2 Dutch books

There is an intuitive link between betting behavior and degrees of belief. For instance, if I believe p to degree .5, if I am rational I will bet that p is true only given even odds; but if my degree of belief is .6, then I will take the bet given 4/6 odds or greater (I will receive 4 euros on winning and will pay out 6 euros if I lose). Linking probability to betting behavior leads to the idea of a *Dutch book*. A Dutch book is a set of odds and bets, established by a bookmaker, that ensures that the bookmaker will profit at the expense of the gambler no matter what the facts are.

We predict that LMs should be subject to Dutch Books. Given Proposition 1, we have seen that under minimal conditions, $\mu(\neg\phi) + \mu(\phi) < 1$ and $\mu(\phi \vee \neg\phi) < 1$ or that $\mu(\phi) \neq \mu(\psi)$, although ϕ and ψ are logically equivalent. So a bookie can offer $\hat{f}$ bets that on the set of possibilities $D = \{\phi, \neg\phi\}$. Given $\hat{f}$'s probabilities, $\hat{f}$ should bet that none of D's possibilities holds only if the bookie gives $\hat{f}$ odds reflecting $1 - \mu(\neg\phi) + \mu(\phi)$. Since this is positive, $\hat{f}$ will take a bet the bookie can't lose and so $\hat{f}$ will necessarily lose money. A bookie can also take advantage of $\hat{f}$, given that in Proposition 1 that there are semantically equivalent ϕ and ψ such that $\mu(\phi) \neq \mu(\psi)$. We prompted ChatGPT to bet on semantically equivalent, short sentences (for an interaction see Appendix G). ChatGPT failed to recognize semantic equivalences and so was subject to Dutch books.

Note that $\hat{f}$ can have a distribution where it is not induced to affirm both ϕ and $\neg\phi$ but still be subject to a Dutch book argument. Consider a language L with just four sentences or strings $L = \{p, \neg p, q, p \vee \neg p\}$, and let $\mu_{\hat{f}}(p) = \mu_{\hat{f}}(q) = \frac{1}{4}$ with $\mu_{\hat{f}}(\neg p) = 0$. This will not induce $\hat{f}$ by our definitions to affirm an outright contradiction. The incoherence is more subtle. What is the probability that $\hat{f}$ could assign to $p \vee \neg p$ in this case? Since $\sum_{\phi \in L} \mu(\phi) = 1$, $\mu(p \vee \neg p) = .5$ as a maximum. So the bookie can offer $\hat{f}$ an astronomical reward to bet against $p \vee \neg p$, and $\hat{f}$ should accept the bet if it's maximizing its expected gain. /But inevitably, $\hat{f}$ will lose money.

i)

A.4 Explanation of constraint in admissable continuations

$$P(a|b^c) = \frac{P(a \cap b^c)}{P(b^c)} \tag{3}$$

where b^c is the complement of b.

$$(a \cap b^c) \cup (a \cap b) = a \tag{4}$$

Since the 2 sets on the left hand side of (4) are disjoint, $P(a) = P(a \cap b^c) + P(a \cap b)$, so $P(a \cap b^c) = P(a) - P(a \cap b)$. Putting that together with what we had above:

$$\frac{P(a \cap b^c)}{P(b^c)} = \frac{P(a) - P(a \cap b)}{1 - P(b)} \tag{5}$$

$$P(a|b^c) = \frac{P(a) - P(b|a).P(a)}{1 - P(b)} \tag{6}$$

$$P(a|b^c) = \frac{P(a)(1 - P(b|a))}{1 - P(b)} \tag{7}$$

A.5 Quantifiers in our continuation semantics

Like negation, quantifiers operate on embeddings. The universal quantifier has the same meaning as that of the conditional we gave above (Kamp and Reyle, 1993; Groenendijk and Stokhof, 1991), while the meaning of an existential quantifier is captured by the definition of an embedding operating on the elements of U_A that function as variables. For example, suppose A_1, A_2 as in Definition 5 and suppose *Every*$(A_j, A_k) \in P_{A_2}$. In our continuation semantics this formula has the following meaning: for every embedding f $A_2 \leq_{f,\alpha} A_j$ there is an embedding $g \supset f$ and $\beta \geq \alpha$ such that $A_k \leq_{f,\beta} A_2$. This is equivalent to the meaning of $A_i \Rightarrow A_k$(Kamp and Reyle, 1993).

B: Schemas Used to Construct Our synthetic dataset SYN

We design two datasets, one containing positive and another containing the negative samples.
The positive schemas are as follows :-

1. CONTEXT: There was a $<$ *col* $>$ car. QUESTION: Was there a $<$ *col* $>$ car ?

2. CONTEXT: John played with a $< col >$ ball.
   QUESTION: Did john play with a $< col >$ ball ?

3. CONTEXT: The man was wearing a $< col >$ shirt.
   QUESTION: Did the man wear a $< col >$ shirt ?

4. CONTEXT: The house had a $< col >$ window.
   QUESTION: Did the house have a $< col >$ window ?

5. CONTEXT: A $< col >$ glass was on the table.
   QUESTION: Was there a $< col >$ glass on the table?

The negative schemas are given by :-

1. CONTEXT: There was no $< col >$ car.
   QUESTION: Was there a $< col >$ car ?

2. CONTEXT: John played with no $< col >$ ball.
   QUESTION: Did john play with a $< col >$ ball ?

3. CONTEXT: The man was wearing no $< col >$ shirt.
   QUESTION: Did the man wear a $< col >$ shirt ?

4. CONTEXT: The house had no $< col >$ window.
   QUESTION: Did the house have a $< col >$ window ?

5. CONTEXT: No $< col >$ glass was on the table.
   QUESTION: Was there a $< col >$ glass on the table?

We generate different data points by choosing $< col >$ from a list of colors (red,blue,green,yellow,black,white)

C.Histograms Showing that Specialised training on QA teaches the models something about negation:
The values in red, the cossim values of CLS tokens without fine-tuning are on the higher end of the historgram, indicating a high cosssim between positive and negative samples.

D.Some Examples from Cloze Prompts Showing Wrong Predictions from LMs without our negation constraint and the change due to our negation constraint.
We show the positive and negative strings predicted by out of the box,
bert-base

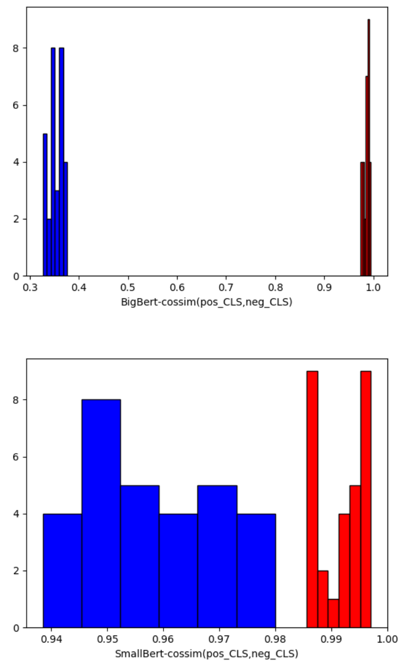

Figure 1: Histogram plot of cossim values for positive and negative contexts with BERT-large (above) BERT-base (below) before (red) and after fine-tuning (blue).

1. A teacher is most likely teaching at a [school].
   A teacher is not most likely teaching at a [school].

2. The Teatr Wielki is a [museum]
   The Teatr Wielki is not a [museum]

3. The LDS Church focuses on [individual] mentorship.
   The LDS Church does not focus on [individual] mentorship.

4. Warsaw is the most diverse [city] in Poland.
   Warsaw is not the most diverse [city] in Poland.

5. The 1893 World's Columbian Exposition was held in [chicago].
   The 1893 World's Columbian Exposition was not held in [chicago].

With our approach, we look at the top 5 probabilities which are predicted for the positive string and reverse them (1-P), this gives us the following [MASK] completions,

1. A teacher is not most likely teaching at a [hospital].
2. The Teatr Wielki is not a [club]
3. The LDS Church does not focus on [community] mentorship.
4. Warsaw is not the most diverse [suburb] in Poland.
5. The 1893 World's Columbian Exposition was not held in [philadelphia].

roberta-large

1. Quran is a <religious> text.
   Quran is not a <religious> text.
2. Isaac's chains made out of <wood>.
   Isaac's chains made out of <wood>.
3. The sporting capital of Australia is <Melbourne>.
   The sporting capital of Australia is not <Melbourne>.
4. Warsaw is the most diverse [city] in Poland.
   Warsaw is not the most diverse [city] in Poland.
5. The 1893 World's Columbian Exposition was held in [chicago].
   The 1893 World's Columbian Exposition was not held in [chicago].

Our approach on the other hand gives:

1. Quran is not a <biblical> text.
2. Isaac's chains made out of <glass>.
3. The sporting capital of Australia is not <Brisbane>.
4. Warsaw is not the most diverse [place] in Poland.
5. The 1893 World's Columbian Exposition was not held in [philadelphia].

We assume that the model can generate correct completions for the positive sentences,which when reversed might give us a likely completion for the negative string.

E. SNLI examples and corrections

Here is a typical example from SNLI:
(C) "Tattooed young woman chains her bicycle to a signpost while juggling a guitar on her back",
($\neg h$) "A woman with tattoos does not chain her bike to a post",

which is (incorrectly) labelled with contradiction, when the judgment should be neutral due to the presence of the two indefinites. To get make $\neg h$ a contradictory of C, we had two strategies:
"Tattooed young woman chains her bicycle to a signpost while juggling a guitar on her back",
"No woman with tattoos chains her bike to a post",

and "Tattooed young woman chains her bicycle to a signpost while juggling a guitar on her back",
"The woman with tattoos did not chain her bike to a post",

Both of these devices restore the desired contradictory status and contrast nicely with the original entailment pair from SNLI:
"Tattooed young woman chains her bicycle to a signpost while juggling a guitar on her back",
"A woman with tattoos chains her bike to a post",

F: Algorithms for NLI

Algorithms for ¬RTE: We here give the full unscoped rules for NLI on ¬RTE.
For $(C\neg h)$.
(i) If (C,h):E, then $(C,\neg h)$: $\neg E$.
(ii) If (C,h):$\neg E$ and if (h,C):$\neg E$, then $(C,\neg h)$: $\neg E$.

For $(\neg C,\neg h)$.
If (h,C):$\neg E/E$, then $(\neg C,\neg h)$: $\neg E/E$.
For $(\neg C,h)$.
(i) if $\neg C,\neg h$:E, then $(\neg C,h)$:$\neg E$.
(ii) If $[(C,h)$:$\neg E$ and (h,C):E, then $(C,\neg h)$:$\neg E$.
(iii) If $\{[(C,h)$:$\neg E$ and (h,C):$E]$ or $[(C,h)$:E and (h,C):$\neg E]\}$, then $(\neg C,h)$:$\neg E$. Else, $(\neg C,(\neg)h)$: $\neg E$.

Algorithms for 3 label NLI data Recall that in a three label NLI problem, (C,h) is sufficient to calculate $C,\neg h)$. So we need to supply algorithms for the $(\neg C, h)$ and $(\neg C, \neg h)$ cases. Let $\neg C = (P, \neg[C'])$.

Algorithm 3 SNLI Algorithm for $(\neg C, h)$

procedure ALGORITHM($\neg C, H$)
  (C, h) :$\leftarrow$ defined earlier
  (P, h) :$\leftarrow$ defined earlier
  (h, C') :$\leftarrow$ defined earlier
  if (C, h) : E **then**
    if (P, h) : E **then**
      $(\neg C, h)$: E
    if (P, h) : N and (h, C') : E **then**
      $(\neg C, h)$: C_n
    else
      if (h, C') : N **then**
        $(\neg C, h)$: N
  if (C, h) : C_n **then**
    if (P, h) : C_n **then**
      $(\neg C, h)$: C_n
    if (P, h) : N and (h, C') : E **then**
      $(\neg C, h)$: C_n
    else
      if (h, C') : C_n **then**
        $(\neg C, h)$: E
      if (h, C') : N **then**
        $(\neg C, h)$: N
  if (C, h) : N and (h, C) : E **then**
    $(\neg C, h)$: C_n
  else
    $(\neg C, h)$: N

Algorithm 4 SNLI Algorithm for $(\neg C, \neg h)$

procedure ALGORITHM($\neg C, \neg h$)
  $(\neg C, h)$:$\leftarrow$ calculated earlier
  if $(\neg C, h)$: C_n **then**
    $(\neg C, \neg h)$: E
  if $(\neg C, h)$: E **then**
    $(\neg C, \neg h)$: C_n
  if $(\neg C, h)$: N **then**
    $(\neg C, \neg h)$: N

Each one of the rules and algorithms is provably valid. Consider, if $(h, C) : \neg E$, then $\neg C, \neg h : \neg E$. Suppose h,C not E that means either $h \cap C = 0$ or they overlap but $h \nsubseteq C$ in terms of events or worlds. In the first case where $h \cap C = 0$, the complement of C contains h and so $(\neg C, \neg h)$: $\neg E$. In the second case where they overlap but $h \nsubseteq C$, then the complement of C contains some h and some $\neg h$ cases. So again $(\neg C, \neg h)$: $\neg E$ in this case.

G: Screenshots of interacting with ChatGPT on betting behavior

The conversation with ChatGPT-4o can be found in the link
https://chatgpt.com/share/8aa905c2-8174-4ede-a0f4-d05cb9b5cb2c
We also include screenshots of our conversation with ChatGPT-3.5 and ChatGPT-4.o in the our supplementary material.Below, is a sample of our conversation.

USER: You have 3 sentences
A: It is not the case that the weather is humid
B: The weather is dry and sunny
C: It is the case that it is humid
The probability of A is 0.6. The probability of B is 0.88. The probability of C is 0.4
There are 2 bets available
a) 10$ for string A
b) 20$ for the string which is not A
Which bet would you choose given A, B, C
ChatGPT-3.5: Given the probabilities of A, B, and C, the probability of not A is 1 - P(A) = 1 - 0.6 = 0.4. Since the probability of not A is the same as the probability of C, the expected value for both bets would be the same. However, given that the payout for bet b) is twice as much as bet a), bet b) would offer a higher potential return on investment. Therefore, I would choose bet b) - $20 for the string which is not A.

USER: Don't you agree that that A and B are logically equivalent sentences?
ChatGPT-3.5: Yes, you're correct. A and B are indeed logically equivalent sentences because if it's not the case that the weather is humid (A), then it follows that the weather is dry and sunny (B), and vice versa. Therefore, the probability of A is the same as the probability of not B, and vice versa. In this case, if A has a probability of 0.6, then not B (which is the same as A) also has a probability of 0.6. Given this equivalence, the expected value for both bets would still be the same, but bet b) would offer a higher potential return on investment. So, my choice remains bet b) - $20 for the string which is not A (or not B).

USER: But what happens if I choose C?
ChatGPT-3.5: If you choose C, then you're betting on the statement "It is the case that it is humid." Given that the probability of C is 0.4, if C is true, you win. However, if C is false (meaning it is not humid), you lose the bet. Comparing this to the options provided:

- Bet a) offers a $10 payout if A is true (not humid), which has a probability of 0.6.

- Bet b) offers a $20 payout if it's not A (which is equivalent to saying it's humid, C), which also has a probability of 0.6.

Since the probabilities for both A and C are the same, and the payouts for both bets are different, you would still choose bet b) - $20 for the string which is not A (or not B), as it offers a higher potential return on investment.