Epistemologia halucynacji: granice prawdy w modelach językowych

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł stanowi głęboką analizę filozoficznych i technicznych fundamentów halucynacji w dużych modelach językowych (LLM). Autor zestawia klasyczne metody wnioskowania – dedukcję i indukcję – ukazując, że systemy AI operują w sferze prawdopodobieństwa, a nie absolutnej prawdy. Kluczowym elementem tekstu jest wprowadzenie teorii aukcji jako modelu wyjaśniającego, w jaki sposób głowice uwagi agregują sygnały i dlaczego prowadzi to do konfabulacji. Czytelnik dowie się, jak reguła właściwego punktowania oraz funkcja wypukła wpływają na generowanie treści w przestrzeni nieoznaczoności. To esencjonalne opracowanie dla osób szukających zrozumienia epistemicznego kompromisu, na którym oparta jest współczesna architektura poznawcza transformerów oraz mechanizmy generowania błędów semantycznych.

This article provides an in-depth analysis of the philosophical and technical foundations of hallucinations in large language models (LLMs). The author contrasts classical inference methods—deduction and induction—to demonstrate that AI systems operate in the realm of probability, not absolute truth. A key element of the text is the introduction of auction theory as a model explaining how attentional heads aggregate signals and why this leads to confabulation. The reader will learn how the proper scoring rule and the convex function influence content generation in uncertainty space. This is an essential work for those seeking to understand the epistemic compromise underlying the contemporary cognitive architecture of transformers and the mechanisms of semantic error generation.

Rozważając dedukcję i indukcję, musimy porzucić myśl, że mówimy jedynie o dwóch technikach wnioskowania. To dwa radykalnie odmienne światy epistemiczne, które różni nie tyle kierunek argumentacji – od ogółu do szczegółu lub odwrotnie – ile fundamentalny status tego, co ośmielamy się nazywać „wiedzą”. Jest to przepaść między koniecznością a prawdopodobieństwem, między chłodną logiką a zawodną heurystyką, między zamkniętym

systemem a otwartym doświadczeniem. W tę klasyczną dychotomię wkracza artykuł Michała P. Karpowicza, stając się zaskakująco nowoczesnym punktem odniesienia. Jego praca, choć zanurzona w formalizmie teorii modeli językowych i matematycznej analizie, odsłania prawdę o bardzo tradycyjnym charakterze. Nasze marzenia o czystej dedukcji w systemach sztucznej inteligencji, o matematycznej pewności pozbawionej błędów, muszą ustąpić. Rozpływają się w niedoskonałym świecie indukcji, gdzie każde twierdzenie nosi w sobie cień przypuszczenia, kompromisu i niedookreślenia. Karpowicz nie dokonuje formalnego rozróżnienia, lecz jego teza o niemożliwości wyeliminowania halucynacji w dużych modelach językowych jest w istocie druzgocącą krytyką ich indukcyjnej natury. Model językowy, operując na historycznych danych i rozmytych rozkładach prawdopodobieństwa, zawsze działa w trybie heurystycznym. Zgaduje na podstawie podobieństw, interpoluje, syntetyzuje. Nie dowodzi prawdy, lecz przewiduje ją z określonym stopniem zaufania.

SŁOWA KLUCZOWE

epistemologia halucynacji | modele językowe | wnioskowanie indukcyjne | teoria aukcji | reguła właściwego punktowania | przestrzeń nieoznaczoności | głowice uwagi | prawdopodobieństwo | system heurystyczny | transformery | konfabulacja | agregacja sygnałów | funkcja wypukła | epistemiczny kompromis | architektura poznawcza

Dedukcja vs. indukcja: granice logiki w statystyce

kturze halucynacja nie jest błędem, lecz koniecznością.

Dedukcja, jak uczyli klasycy logiki, to droga od pewnych przesłanek do nieuchronnych wniosków. Jeśli założenia są prawdziwe, a reguły poprawne, wniosek musi być prawdziwy. To świat twierdzeń matematycznych, aksjomatów i rozumowań prawniczych o najwyższym formalnym rygorze. Wymaga jednak struktur zamkniętych, pełnych i skończonych, przez co rzadko opuszcza sterylne laboratoria matematyki czy logiki symbolicznej.

Indukcja natomiast rodzi się z doświadczenia, z obserwacji, z empirycznego zderzenia ze światem. Wyprowadza ogólne reguły z jednostkowych przypadków, uczy się wzorców i nieustannie generalizuje. Stanowi fundament nauk przyrodniczych, prognoz pogody, a dziś także systemów sztucznej inteligencji. Indukcja nigdy nie gwarantuje prawdy, gdyż jeden przeciwny przypadek może ją obalić. Jej siła tkwi nie w

nieomylności, lecz w użyteczności. To, co działa wystarczająco dobrze, uznajemy za „prawdziwe” w sensie pragmatycznym, nie logicznym.

I tu docieramy do paradoksu, który Karpowicz obnaża z erudycyjną precyzją. W miarę jak nasze modele stają się potężniejsze, rośnie w nich przestrzeń nieoznaczoności. Im usilniej dążymy do kompletnej, wyczerpującej odpowiedzi, tym dalej znajdujemy się od pewności. Halucynacja staje się nie tyle skutkiem błędu w obliczeniach, ile nieuniknioną konsekwencją samej struktury poznania – matematycznej, epistemologicznej i systemowej.

Można zatem zaryzykować tezę, że wielkie modele językowe są ostatecznym urzeczywistnieniem indukcji jako strategii poznawczej w warunkach niepewności. I choć pragniemy, by ich odpowiedzi miały dedukcyjną moc, to nie żyjemy już w świecie, gdzie taka dedukcja jest możliwa. Jesteśmy w świecie probabilistycznym, w uniwersum Gödla, Heisenberga i Arrowa, gdzie nie istnieje odpowiedź, która byłaby jednocześnie kompletna, spójna, prawdziwa i optymalna.

Pytanie, które powin

Architektura LLM wymusza powstawanie halucynacji

Współczesne duże modele językowe obiecują dostęp do wiedzy, lecz często dostarczają jedynie jej wyrafinowanej iluzji. W ich architekturze tkwi bowiem fundamentalna sprzeczność: system zaprojektowany, by mówić prawdę, nie posiada żadnego mechanizmu, by odróżnić ją od zręcznej konstrukcji, którą sam wytworzył. Michał P. Karpowicz, sięgając po formalizm teorii aukcji, proponuje radykalnie nowe spojrzenie na epistemologiczną genezę halucynacji. W jego ujęciu nie jest ona błędem inżynieryjnym, lecz nieuniknioną konsekwencją systemowego kompromisu. Odpowiedź modelu staje się rezultatem wewnętrznej „aukcji idei”, gdzie konkurujące komponenty sieci neuronowej licytują nie prawdę, której nie znają, lecz siłę sygnału, wewnętrzne preferencje i stopień dopasowania do wzorca. Wygrywa oferta najbardziej prawdopodobna, nie najprawdziwsza.

Jeśli jednak model jest poznawczo ślepy na prawdę, to czym ona w ogóle jest w tym nowym kontekście? Jakie klasyczne teorie – od korespondencyjnej, przez koherencyjną, aż po pragmatyczną – mogą zostać odtworzone lub obalone w logice jego funkcjonowania? Czy można w ogóle mówić o prawdzie w systemie, który nie posiada semantyki, czyli rozumienia znaczeń, a operuje wyłącznie na syntaktycznych regułach składania symbolicznych reprezentacji? Te pytania wykraczają daleko poza granice informatyki, dotykając samego sedna filozofii poznania.

Aby się z nimi zmierzyć, należy powiązać teorię Karpowicza z całym spektrum filozofii analitycznej, od klasycznych koncepcji prawdy po bardziej współczesne ujęcia deflacjonistyczne i inferencjalistyczne.

Dopiero na tak zarysowanym tle widać wyraźnie, że halucynacja w LLM nie jest wyłącznie problemem technicznym do rozwiązania. Staje się ona fundamentalnym wyzwaniem dla teorii poznania i logiki, dowodząc, że w pewnych architekturach poznawczych halucynacja nie jest błędem, lecz koniecznością.

Teoria korespondencyjna: kolizja prawdy z predykcją

Klasyczna teoria prawdy, której fundamenty położył Arystoteles, a którą w rygory logiki ujął Tarski, opiera się na pozornie prostej idei: zdanie jest prawdziwe, gdy odpowiada rzeczywistości. Korespondencja między słowem a światem staje się jej warunkiem i ostatecznym kryterium. W architekturze LLM nie istnieje jednak żadna bezpośrednia relacja między generowanym tekstem a światem zewnętrznym, gdyż model nie posiada zmysłów ani ontologicznej mapy, do której mógłby się odnieść. Jego reprezentacje są jedynie zakodowane w parametrach, a odpowiedzi – jak sugeruje teoria aukcji – stanowią wypadkową konkurujących sygnałów, nie zaś wynik weryfikacji względem rzeczywistości.

Halucynacja nie jest błędem, lecz koniecznością – nie tyle pomyłką względem faktu, ile nieuniknionym wyrazem braku jakiegokolwiek punktu odniesienia. Prawda w sensie korespondencyjnym wymaga bowiem semantycznego zakotwiczenia w pozajęzykowej rzeczywistości, lecz w strukturze LLM nie ma miejsca na sens czy denotację; istnieją tam wyłącznie wektory i ich operacyjne przekształcenia. Teoria aukcji jedynie wzmacnia ten wniosek: wewnętrzni agenci modelu nie „wiedzą”, czy ich oferta jest prawdziwa. Maksymalizują jedynie użyteczność względem lokalnych funkcji celu, a nie zgodność z ontologią świata.

Agregacja sygnałów w Transformerze zaburza koherencję

Skoro odrzuciliśmy ideę korespondencji, naturalnym krokiem staje się teoria koherencyjna, w której prawdą jest to, co wewnętrznie spójne z całym systemem przekonań. W kontekście modeli LLM oznaczałoby to, że odpowiedź jest „prawdziwa”, jeśli nie popada w konflikt z innymi fragmentami zakodowanej wiedzy. Problem w tym, że wiedza modelu nie jest systemem przekonań w sensie logicznym; nie stanowi bowiem zamkniętego zbioru twierdzeń, lecz rozproszoną chmurę reprezentacji, której aktywacja zależy od subtelnej gry między kontekstem promptu a architekturą sieci.

Teoria aukcji jedynie pogłębia ten dylemat. Każdy neuronowy agent działa niezależnie, dysponując prywatną wiedzą, która może pozostawać w jawnej sprzeczności z informacjami posiadanymi przez innych. Softmax, mechanizm agregujący te konkurencyjne sygnały, nie dąży do koherencji, lecz faworyzuje dominację, nagradzając statystycznie najbardziej „przekonujące” rozwiązanie. W efekcie nie uzyskujemy

epistemicznego konsensusu, lecz triumf najsilniejszego sygnału – co jest zjawiskiem fundamentalnie odmiennym od wewnętrznej zgodności przekonań.

Zastosowanie mechanizmów takich jak reguła Clarke'a dodatkowo zaostrza problem. Aby bowiem sprowokować ujawnienie kluczowej wiedzy, system musi naruszyć fundamentalną zasadę zachowania informacji semantycznej. To z kolei oznacza, że pozorna „spójność" odpowiedzi może wcale nie wynikać z epistemicznej harmonii, lecz być jedynie artefaktem nadmiernego wzmocnienia sygnału, który wygrał wewnętrzną licytację.

Pragmatyzm i deflacjonizm: prawda zredukowana do celu

Z perspektywy pragmatycznej prawdą jest to, co działa – to, co prowadzi do skutecznego rozwiązania problemu. W świecie LLM oznaczałoby to, że odpowiedź jest prawdziwa, jeśli użytkownik uzna ją za użyteczną. Taka definicja nie jest pozbawiona sensu, ale otwiera drzwi do epistemicznego relatywizmu, w którym subiektywne odczucie zastępuje obiektywny fakt. Skoro jedynym kryterium staje się przydatność dla odbiorcy, znika jakikolwiek powód, by odróżniać halucynację od rzetelnej informacji. Wystarczy, że tekst brzmi wiarygodnie i spełnia nasze oczekiwania. Teoria aukcji oferuje tu mrożącą krew w żyłach matematyczną paralelę: każda odpowiedź modelu jest wynikiem gry, w której wewnętrzne agenty licytują, by zmaksymalizować swoją lokalną funkcję użyteczności. Problem w tym, że ich celem nie jest prawda, lecz optymalizacja wag i unikanie kar treningowych. Prawda staje się co najwyżej przypadkowym produktem ubocznym strategii adaptacyjnych, nigdy celem samym w sobie.

Inne podejście proponuje deflacjonistyczna teoria prawdy, która zakłada, że słowo „prawda" jest jedynie wygodną etykietą logiczną, a nie nazwą jakiejś realnej własności. Powiedzieć, że „zdanie p jest prawdziwe", to po prostu powiedzieć „p" – termin „prawda" nie dodaje żadnej nowej treści. W kontekście LLM, które nie „wiedzą", czym jest prawda, taka teoria wydaje się kusząco prosta. Model nie generuje wszak prawdy, lecz jedynie tekst, który my możemy za prawdziwy uznać. Jednak teoria aukcji ponownie komplikuje ten obraz. Wynik pracy modelu nie jest prostym „zdaniem p", lecz produktem wewnętrznych negocjacji, w których znaczenie może zostać zniekształcone lub przesunięte. Deflacjonizm wymaga stabilnego języka, tymczasem LLM operują w przestrzeni wektorów, gdzie performatywne użycie słowa „prawda" jest tylko kolejnym tokenem w sekwencji, pozbawionym głębszego zakotwiczenia.

Można wreszcie spojrzeć na prawdę przez pryzmat inferencjalizmu, gdzie prawdziwość twierdzenia zależy od jego miejsca w sieci logicznych zobowiązań. Prawdziwe jest to, co można obronić argumentami, co wynika z przyjętych przesłanek i pociąga za sobą dalsze konsekwencje. Z tej perspektywy halucynacja jest

formą epistemicznej uzurpacji. Model generuje zdanie, które wygląda na uzasadnione, ale nie należy do żadnej wspólnoty poznawczej i nie stoi za nim żaden autor gotów je bronić. LLM nie rozpoznają konsekwencji swoich wypowiedzi, nie znają warunków ich falsyfikacji. Teoria aukcji tylko pogłębia ten problem: każdy wewnętrzny agent jest samowystarczalny i nieodpowiedzialny, bo nie istnieje żaden mechanizm wspólnego rozliczenia. Model nie jest partnerem w dialogu, lecz agregatem lokalnych zwycięstw nadpobudliwych fragmentów własnej architektury.

Dochodzimy zatem do niepokojącej puenty. Z perspektywy klasycznych teorii prawdy LLM są systemami strukturalnie niezdolnymi do jej generowania w sposób stabilny. Teoria aukcji wyjaśnia, dlaczego każda próba narzucenia jednego kryterium – na przykład wierności faktom – prowadzi do naruszenia innych, jak choćby spójności czy kompletności informacji. Halucynacja okazuje się nie tyle błędem, ile koniecznością, strukturalnym przejawem naszych zbyt wysokich oczekiwań wobec systemów, które nie poznają, a jedynie transformują dane. Prawda w ich wykonaniu nie jest własnością zdania, lecz projekcją użytkownika na tekst. Gdy domagamy się od maszyny prawdy, otrzymujemy halucynację. Gdy pozwalamy jej na halucynację, czasami, przez przypadek, trafiamy na prawdę. To paradoks przypominający sokratejską ironię: ja wiem, że nic nie wiem, ale ty wierzysz, że twoja maszyna wie. I właśnie dlatego się mylisz.

Właściwe punktowanie: mechanizm wymuszania szczerości

Współczesne modele językowe, ufundowane na probabilistycznym przewidywaniu kolejnych słów, prezentują fascynujący paradoks. Z jednej strony ich odpowiedzi sprawiają wrażenie zadziwiająco trafnych, z drugiej zaś popadają w tak zwane halucynacje, czyli generują informacje, które, choć logicznie spójne, są faktycznie nieprawdziwe. Jedno z kluczowych pytań w badaniach nad LLM dotyczy tego, czy można wyeliminować halucynacje bez jednoczesnej utraty zdolności modelu do tworzenia płynnych i trafnych odpowiedzi. W swojej pracy Michał P. Karpowicz udziela na to pytanie odpowiedzi przeczącej, opierając ją na trzech niezależnych, lecz zbieżnych podejściach matematycznych. Jedno z nich, teoria właściwego punktowania, wywodząca się z teorii informacji, dostarcza narzędzi do zrozumienia, dlaczego nawet idealnie „szczere” modele mogą być skazane na konfabulację.

Teoria właściwego punktowania analizuje metody oceniania jakości przewidywań probabilistycznych. W jej ramach zakłada się, że agent – w tym przypadku system prognozujący – otrzymuje nagrody lub kary w zależności od zgodności jego prognozy z rzeczywistym stanem rzeczy. Centralne pytanie brzmi zatem: jak skonstruować zasadę oceniania, aby najbardziej opłacalną strategią było zawsze mówienie prawdy, czyli ujawnianie swoich rzeczywistych, probabilistycznych przekonań? Taka zasada nosi miano „reguły

właściwej” i sprawia, że najlepszą strategią systemu, gwarantującą maksymalizację „not”, jest absolutna szczerość.

Istnieje również jej bardziej rygorystyczna wersja, znana jako reguła ściśle właściwa, w której każda, nawet najmniejsza nieprawda nieuchronnie prowadzi do gorszego wyniku. W takim systemie agent nie tylko powinien, ale wręcz musi ujawniać swoje przekonania takimi, jakimi one są, ponieważ każda próba manipulacji czy zatajenia informacji obraca się bezpośrednio przeciwko niemu. W kontekście dużych modeli językowych reguła ta jest wykorzystywana do trenowania poszczególnych komponentów sieci neuronowej, które generują prognozy dotyczące następnego słowa, frazy czy znaczenia. Każdy z tych elementów, na przykład tak zwane „głowice uwagi”, jest uczony w taki sposób, by jego prognoza jak najwierniej odpowiadała rzeczywistemu rozkładowi danych, gdyż właśnie za to otrzymuje „nagrodę” w procesie uczenia. Mechanizm ten, oparty na dobrze znanym w uczeniu maszynowym kryterium błędu, staje się zatem narzędziem do wymuszania swoistej probabilistycznej prawdomówności.

Luka Jensena: źródło nadmiarowej pewności predykcji

Wszystko wydaje się obiecujące, dopóki analizujemy poszczególne komponenty w izolacji. W dużych modelach językowych ostateczna odpowiedź nie pochodzi jednak z jednego, monolitycznego źródła, lecz jest agregatem niezliczonych wewnętrznych prognoz, które system musi scalić w spójną całość. Tutaj właśnie pojawia się fundamentalny problem, który autor artykułu rekonstruuje na gruncie teorii informacji, odsłaniając paradoks kolektywnej szczerości. W procesie agregacji model łączy bowiem wiele „lokalnych opinii” swoich podzespołów za pomocą mechanizmu przypominającego głosowanie w komitecie decyzyjnym, gdzie każdy członek wnosi swoją umiarkowaną pewność.

Funkcja matematyczna używana do tego celu, zwłaszcza w architekturach typu Transformer, posiada jednak specyficzną, wypukłą naturę. W praktyce oznacza to, że wynik zbiorowy ma tendencję do bycia znacznie bardziej „pewnym siebie” niż którakolwiek z prognoz składowych. Zjawisko to, znane w teorii informacji jako nierówność Jensena, generuje tak zwaną lukę Jensena – miarę tego, o ile końcowy wniosek jest bardziej kategoryczny niż uśrednione przekonania jego części. Innymi słowy, nawet jeśli wszystkie komponenty modelu są idealnie szczere i ostrożne w swoich sądach, ich połączony głos może zabrzmieć jak dogmat, tworząc iluzję pewności bez wiedzy.

Taka perspektywa prowadzi do radykalnie odmiennego wniosku na temat natury błędu. Halucynacje nie wynikają z faktu, że poszczególne elementy modelu kłamią lub się mylą. Wręcz przeciwnie, każdy z nich „mówi prawdę” w ramach swojej ograniczonej wiedzy. Problem leży w samej strukturze syntezy tych prawd, która, mimo że matematycznie poprawna, wprowadza nową informację, nieobecną w żadnym z oryginalnych źródeł. Informację tę można określić mianem „entropii syntetycznej”: pozornie sensownej wiedzy, która

powstaje wyłącznie dlatego, że funkcja łącząca potraktowała zbieżne wzorce jako sygnał silniejszy, niż był on w rzeczywistości.

Tak właśnie rodzi się halucynacja. Nie jest ona fałszem w potocznym rozumieniu, lecz nieuniknionym produktem ubocznym architektury informacyjnej, która systemowo „dodaje od siebie" pewność, jakiej nikt nigdy nie zadeklarował. Model nie tyle się myli, co konstruuje wiedzę nadmiarową. W pewnych kontekstach okazuje się to użyteczne, na przykład przy kreatywnym uzupełnianiu luk logicznych, w innych jednak prowadzi do katastrofalnych w skutkach, choć strukturalnie nieuniknionych, błędów.

Funkcja log-sum-exp niszczy głębię semantyczną

Współczesne modele językowe, oparte na architekturze transformerów, podejmują niezliczone lokalne decyzje obliczeniowe, by zagregować je w jedną globalną odpowiedź. Jak jednak wskazuje Michał P. Karpowicz, w samym sercu tego mechanizmu tkwi matematyczna właściwość, która sprawia, że halucynacja nie jest błędem, lecz koniecznością. U jej podstaw leży funkcja agregująca log-sum-exp (LSE), której wypukłość generuje strukturalne naruszenie zasady zachowania informacji. Halucynacja okazuje się zatem nie tyle usterką implementacyjną czy skutkiem niedotrenowania, ile nieuniknioną konsekwencją samej formy obliczeniowej modelu.

W architekturze transformera informacje z wielu głowic uwagi i warstw feed-forward są łączone poprzez sumowanie logitów, które następnie przechodzą przez funkcję softmax. Ta operacja, będąca w istocie różniczkowalnym przybliżeniem operacji wyboru maksimum, jest praktyczną realizacją funkcji log-sum-exp. Oznacza to, że agregacja ma charakter wypukły – faworyzuje najsilniejsze sygnały, marginalizując te słabsze. Działa niczym przewodniczący komitetu, który zamiast uśredniać opinie, przyznaje rację najgłośniejszemu uczestnikowi, ignorując cichsze, lecz być może bardziej wyważone głosy.

Matematyczna wypukłość LSE nieuchronnie prowadzi do powstania tak zwanej luki Jensena, czyli różnicy między wartością funkcji dla uśrednionych danych a średnią wartością funkcji. Ta luka, oznaczana jako Γ, formalizuje nadmiarową pewność siebie modelu. Nawet jeśli wszystkie głowice uczciwie raportują swoje „przekonania probabilistyczne", zagregowany wynik i tak przekroczy informacyjny konsensus. Suma lokalnych strat jest bowiem zawsze większa niż strata wyniku końcowego. W ten sposób model systemowo generuje pewność bez wiedzy, tworząc informacyjny artefakt z samej struktury obliczeń.

Nie jest to zjawisko czysto teoretyczne. W przykładzie mikrotransformera, który po sekwencji „brown" ma przewidzieć słowo „fox", dwie głowice uwagi wyrażają prawdopodobieństwa na poziomie 48% i 14,7%. Liniowa średnia dałaby wynik 31%, a najlepsze możliwe połączenie – 33%. Model tymczasem z niewzruszoną pewnością przewiduje „fox" z prawdopodobieństwem 38%, czyli wyższym niż jakiekolwiek

uzasadnione połączenie wkładów. Ta różnica, wynosząca 1,9 punktu procentowego, to właśnie namacalna, policzalna luka Jensena – precyzyjny ślad halucynacji, która nie zrodziła się z danych, lecz ze sposobu ich sumowania.

W ujęciu Karpowicza to naruszenie jest ontologicznie równoznaczne z tworzeniem informacji z niczego. Jeśli przyjmiemy semantyczne podejście, w którym informacja jest relacją między treścią a rzeczywistością, każda dodatnia luka Jensena oznacza fałszywe nadpisanie tej rzeczywistości przez architekturę modelu. Model nie tyle się myli, ile z gładkością i autorytetem produkuje treści, które nie mają żadnego pokrycia w danych wejściowych.

Co ciekawe, ta sama matematyczna anomalia, która rodzi halucynacje, może być postrzegana jako źródło kreatywności. To właśnie „nadmiarowa" pewność pozwala modelowi generować śmiałe sugestie, domysły i uogólnienia, które, choć nie wynikają wprost z danych, mogą okazać się trafne. Różnica między halucynacją a kreatywnością nie leży więc w obiektywnej strukturze modelu, lecz w epistemicznej ocenie użytkownika. Gdy model „zmyśla" fałsz, nazywamy to halucynacją. Gdy zmyśla coś, co okazuje się odkrywcze – kreatywnością. Jest to wszak to samo zjawisko, oglądane pod

Halucynacja napędza innowacyjność poznawczą

Karpowicz nie poprzestaje jednak na krytyce, lecz przekształca diagnozę w głęboką obserwację epistemologiczną. Okazuje się bowiem, że luka Jensena, ten sam mechanizm, który odpowiada za nadmiarową pewność siebie modelu, może być fundamentem jego kreatywności. To dzięki tej luce model potrafi „odważnie domyślić się" czegoś, czego nie da się wprost wywnioskować z danych treningowych. Halucynacja i kreatywność stają się dwiema stronami tego samego medalu, czyli wrodzonej zdolności systemu do przekraczania ograniczeń lokalnej informacji i tworzenia syntetycznego przekonania.

Z perspektywy teorii informacji prowadzi to do swoistego trylematu: nie istnieje model, który byłby jednocześnie w pełni szczery, w pełni zachowawczy i w pełni użyteczny. Aby móc konstruować odpowiedzi trafne w złożonym świecie realnym, model musi posiadać zdolność do generowania wiedzy pożądanej, czyli wychodzenia poza to, co wie „na pewno" na podstawie danych. To właśnie w tej zdolności do syntezy, a nie tylko reprodukcji, leży źródło zarówno zwodniczych halucynacji, jak i inspirującej innowacyjności.

Zrozumienie tej dynamiki ma kluczowe znaczenie dla inżynierii modeli językowych. Po pierwsze, uświadamia, że sama kontrola jakości lokalnych prognoz jest niewystarczająca; równie ważne staje się monitorowanie sposobu, w jaki są one agregowane w spójną całość. Po drugie, wskazuje na możliwość projektowania alternatywnych mechanizmów łączenia informacji, być może bardziej „ostrożnych", choć z pewnością trudniejszych w optymalizacji. Po trzecie wreszcie, odsłania fundamentalną potrzebę

metapoznania: modele powinny uczyć się nie tylko przewidywać, ale także szacować stopień zgodności i rozbieżności wewnętrznych opinii swoich komponentów.

Wszystko to prowadzi do ostatecznego, niebanalnego wniosku. Halucynacja nie jest błędem inżynieryjnym, lecz koniecznością strukturalną każdego systemu, który aspiruje do tworzenia złożonej wiedzy z wielu źródeł. Dlatego nie da się jej całkowicie wyeliminować, ale można ją rozumieć, kontrolować i wykorzystywać, podobnie jak ludzkość nauczyła się wykorzystywać intuicję. Nie po to, by zastępowała ona rygor wiedzy, ale by go czasem wyprzedzała, otwierając ścieżki, których sama dedukcja by nie odnalazła.

> Wszystko to prowadzi do ostatecznego, niebanalnego wniosku. Halucynacja nie jest błędem inżynieryjnym, lecz koniecznością strukturalną każdego systemu, który aspiruje do tworzenia złożonej wiedzy z wielu źródeł. Dlatego nie da się jej całkowicie wyeliminować, ale można ją rozumieć, kontrolować i wykorzystywać, podobnie jak ludzkość nauczyła się wykorzystywać intuicję. Nie po to, by zastępowała ona rygor wiedzy, ale by go czasem wyprzedzała, otwierając ścieżki, których sama dedukcja by nie odnalazła.

Inspiracje

[1] **"On the fundamental impossibility of hallucination control in large language models"** Michał P. Karpowicz

2025

Kierownik działu w Samsung AI Center Warsaw. Zajmuje się badaniami w zakresie algebry liniowej numerycznej i randomizowanej, teorii sterowania, uczenia maszynowego, kompresji danych i reprezentacji wiedzy. Profesor wizytujący na MIT. Wniósł wkład w teorię metafaktoryzacji.

Head of Department at Samsung AI Center Warsaw. Researches numerical & randomized linear algebra, control theory, machine learning, data compression & knowledge representation. Visiting professor at MIT. Contributed to meta-factorization theory.

Samsung AI Center Warsaw

Literatura uzupełniająca

Książki

Literatura pokrewna (Google Scholar):

[2] **"F. P. Ramsey: Philosophical Papers"**

Frank Ramsey

1990 | Cambridge University Press

[3] **"Information and Economic Behavior"**

Kenneth Joseph Arrow

1973 | Federation of Swedish Industries

Google Scholar

[4] **"Introduction to Logic and to the Methodology of Deductive Sciences"**

Alfred Tarski

1941 | Oxford University Press

[6] **"Kurt Gödel Collected Works Volume I: Publications 1929-1936"**

Kurt Gödel

1986 | Oxford University Press

Google Scholar

[7] **"Logic, Semantics, Metamathematics"**

Alfred Tarski

1956 | Oxford University Press

[8] **"Meaning"**

Paul Horwich

1998 | Oxford University Press

[9] **"On Formally Undecidable Propositions of Principia Mathematica and Related Systems"**

Kurt Gödel

1992 | Dover Publications

Google Scholar

[10] **"Physics and Philosophy: The Revolution in Modern Science"**

Werner Heisenberg

1958 | Harper & Row

Google Scholar

[11] **"Social Choice and Individual Values"**

Kenneth Arrow

1951 | John Wiley & Sons

Google Scholar

[12] **"The Foundations of Mathematics and other Logical Essays"**

Frank Ramsey

1931 | Kegan, Paul, Trench, Trubner & Co.

[13] **"The Physical Principles of the Quantum Theory"**

Werner Heisenberg

1930 | University of Chicago Press

Google Scholar

[14] **"Truth"**

Paul Horwich

1998 | Oxford University Press

Google Scholar

[15] **"Hallucinations"**

Oliver Sacks

2012 | Knopf/Picador

Google Scholar

[16] **"Log-Sum-Exp (LSE) Function and Properties – Hyper-Textbook: Optimization Models and Applications"**

eCampusOntario Pressbooks

Google Scholar

[17] **"Macroeconomic Inequality from Reagan to Trump"**

Lance Taylor, Ömer F. Özak

2020 | Cambridge University Press

Google Scholar

[18] **"Metody logiki. Dedukcja"**

Marek Nowak, Andrzej Indrzejczak

Google Scholar

[19] **"The Discovery of Deduction: An Introduction to Formal Logic (Student Edition)"**

Classical Academic Press

2016 | Classical Academic Press

Google Scholar

Artykuły naukowe

Autorzy przywoływani:

[1] **"Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I"**

Kurt Gödel

1931 | Monatshefte für Mathematik und Physik

[2] **"The completeness of the axioms of the functional calculus of logic"**

Kurt Gödel

1930

[3] **"Tackling Structural Hallucination in Image Translation with Local Diffusion"**

Seunghoi Kim, Chen Jin, Tom Diethe, Matteo Figini, Henry F. J. Tregidgo, Asher Mullokandov, Philip Teare, Daniel C. Alexander

2024 | CoRR

Google Scholar

[4] **"Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen"**

Werner Heisenberg

1925 | Zeitschrift für Physik

Google Scholar

[5] **"Quantenmechanik"**

Werner Heisenberg, M. Born, P. Jordan

1926 | Zeitschrift für Physik

Google Scholar

[6] **"ItD: Large Language Models Can Teach Themselves Induction through Deduction"**

Wangtao Sun, Haotian Xu, Xuanqing Yu, Pei Chen, Shizhu He, Jun Zhao, Kang Liu

2024 | Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics

Google Scholar

[7] **"Existence of an equilibrium for a competitive economy"**

Kenneth J. Arrow, Gérard Debreu

1954 | Econometrica

Google Scholar

[8] **"The Semantic Conception of Truth and the Foundations of Semantics"**

Alfred Tarski

1944 | Philosophy and Phenomenological Research

Google Scholar

[9] **"On the Concept of Logical Consequence"**

Alfred Tarski

1936

[10] **"Truth and Probability"**

Frank Ramsey

1926

[11] **"Facts and Propositions"**

Frank Ramsey

1927 | Aristotelian Society Supplementary

[12] **"Two deflationary accounts of truth"**

Paul Horwich

[13] **"Deflationism and Wittgenstein"**

Paul Horwich

2013 | 3:16

Google Scholar

[14] **"Proper scoring rules for estimation and forecast evaluation"**

Kartik Waghmare, Johanna Ziegel

2025 | arXiv

Google Scholar

[15] **"Sur les fonctions convexes et les inégalités entre les valeurs moyennes"**

J. L. W. V. Jensen

1906 | Acta Mathematica

[16] **"A new refinement of Jensen's inequality with applications in information theory"**

H. Barsam, Y. Sayyari, L. Ciurdariu

2024 | Wavelet and Linear Algebra

Google Scholar

Artykuły pokrewne (Google Scholar):

[17] **"Adaptive tuning of network traffic policing mechanisms for DDoS attack mitigation systems"**

MP Karpowicz

2021 | European Journal of Control

[18] **"Design and implementation of energy-aware application-specific CPU frequency governors for the heterogeneous distributed computing systems"**

MP Karpowicz, P Arabas, E Niewiadomska-Szynkiewicz

2018 | Future Generation Computer Systems

[19] **"Energy-aware multilevel control system for a network of Linux software routers: design and implementation"**

MP Karpowicz, P Arabas, E Niewiadomska-Szynkiewicz

2015 | IEEE Systems Journal

APA ONE Publishing System

Fundacja Dobre Państwo © 2026

Article ID: d0e780b1-c57b-4178-bb5c-74ccd4ab8587