

Płynne pojęcia i poszarpana inteligencja: modele językowe w świetle teorii Douglasa Hofstadtera



AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje funkcjonalną zbieżność współczesnych modeli językowych (LLM) z teorią FARG Douglasa Hofstadtera. Autor wskazuje, że choć LLM nie korzystają z jawnej architektury kodletów, ich zdolność do rozwiązywania zadań analogicznych i interpretacji metafor sugeruje podobną elastyczność poznawczą. Kluczowym elementem jest mechanizm attention w transformerach, który pozwala na dynamiczne tworzenie reprezentacji kontekstowych, odrzucając sztywny model symboliczny. Choć LLM opierają się na statystycznym przewidywaniu tokenów, skala ich treningu wymusza powstawanie wysokopoziomowych struktur abstrakcyjnych. Tekst podkreśla, że współczesne AI nie mieści się w prostym podziale na systemy symboliczne i koneksjonistyczne, oferując nową formę płynności semantycznej, która funkcjonalnie przypomina ludzkie procesy poznawcze.

The article analyzes the functional convergence of contemporary Large Language Models (LLMs) with Douglas Hofstadter's FARG theory. The author indicates that although LLMs do not use an explicit copycat architecture, their ability to solve analogical tasks and interpret metaphors suggests similar cognitive flexibility. A key element is the attention mechanism in transformers, which allows for the dynamic creation of contextual representations, rejecting a rigid symbolic model. Although LLMs rely on statistical token prediction, the scale of their training forces the emergence of high-level abstract structures. The text emphasizes that modern AI does not fit into a simple division between symbolic and connectionist systems, offering a new form of semantic fluidity that functionally resembles human cognitive processes.

Artykuł analizuje relację między współczesnymi wielkimi modelami językowymi (LLM) a koncepcją płynnych pojęć Douglasa Hofstadtera i jego architekturą FARG. Autor stawia tezę, że choć LLM nie implementują sztywnych struktur FARG, wykazują z nimi funkcjonalną zbieżność poprzez zdolność do dynamicznej rekonfiguracji reprezentacji w zależności od kontekstu. Tekst podważa binarny spór o to, czy AI „rozumie” treści, czy jedynie statystycznie je przetwarza, sugerując, że statystyczna optymalizacja może prowadzić do wyłonienia się wysokopoziomowych abstrakcji i kompetencji analogicznych. Kluczowym punktem analizy jest zjawisko „poszarpanej inteligencji” (jagged intelligence) – paradoksalnej niespójności, w której modele rozwiązują ekstremalnie trudne problemy, jednocześnie zawodząc w zadaniach banalnych dla człowieka. W konsekwencji autor postuluje przejście od prostych testów behawioralnych ku metodologii „eksperymentów rozproszonych”, które badają stabilność generalizacji strukturalnej i odporność systemu na perturbacje powierzchniowe, co pozwala realnie ocenić głębokość poznawczą AI bez popadania w antropomorfizmy.

SŁOWA KLUCZOWE

modele językowe

teoria Douglasa Hofstadtera

architektura FARG

płynne pojęcia

poszarpana inteligencja

mechanizm attention

interpretowalność mechanistyczna

in-context learning

odporność semantyczna

reprezentacje rozproszone

global workspace

zdolność analogiczna

superpozycja w sieciach neuronowych

jagged frontier

inwariancja

Funkcjonalna zbieżność LLM i architektury FARG

Pojawienie się wielkich modeli językowych zmienia sposób, w jaki należy czytać program FARG. Jeszcze w latach dziewięćdziesiątych pytanie mogło brzmieć: czy aby uzyskać inteligencję podobną do ludzkiej trzeba jawnie skonstruować architekturę płynnych pojęć, w której małe procesy podpoznawcze konkurują o zasoby, reprezentacje przebudowują się pod wpływem kontekstu, a analogia powstaje dzięki dynamicznym poślizgom pomiędzy kategoriami? Współczesne systemy wymuszają inną formę tego pytania.

Najbardziej kompetentne modele językowe nie zostały zbudowane według schematu Slipnet-Workspace-Coderack, nie operują na kodletach Hofstadtera i nie posiadają ręcznie skonstruowanej

sieci pojęć. A mimo to potrafią rozwiązywać zadania analogiczne, tłumaczyć między językami, generować kod, interpretować metafory, rekonstruować wzorce, przekształcać styl oraz dostosowywać zachowanie do kontekstu. Fakt ten nie rozstrzyga sporu o naturę inteligencji, lecz uniemożliwia traktowanie konkretnej architektury FARG jako jedynej możliwej drogi do funkcjonalnej elastyczności.

Autor książki przeciwstawia tradycyjną sztuczną inteligencję symboliczną podejściu FARG przede wszystkim na poziomie reprezentacji. W klasycznym obrazie symbolicznym reprezentacja ma być względnie jawna, stabilna i poddawana operacjom według określonych reguł; w FARG struktury pojęciowe pozostają płynne, kontekstowe i częściowo wyłaniają się z lokalnych procesów. Współczesny model językowy nie mieści się jednak wygodnie po żadnej stronie tej opozycji. Nie jest klasycznym systemem symbolicznym, ponieważ jego wiedza nie występuje jako łatwo odczytywalny zbiór zdań, reguł i pojęć zapisanych w jawnej bazie. Nie jest też FARG-owskim systemem kodletów – jego kompetencje wynikają z uczenia parametrów wielowarstwowej sieci neuronowej na podstawie ogromnej liczby przykładów, a reprezentacje mają charakter rozproszony i wysokowymiarowy.

Kluczowym momentem technologicznym było wprowadzenie w 2017 roku architektury transformera. Zastąpiła ona konieczność sekwencyjnego przetwarzania, znaną z sieci rekurencyjnych, mechanizmami attention, które pozwalają modelowi dynamicznie ważyć relacje pomiędzy elementami sekwencji. Choć pierwotna publikacja Vaswaniego i współautorów dotyczyła przede wszystkim translacji maszynowej, architektura ta okazała się niezwykle podatna na skalowanie i adaptację do innych zadań. Sam ten fakt nie mówi jeszcze niczego o „rozumieniu”, pokazuje jednak zmianę architektoniczną istotną z punktu widzenia filozofii FARG: wpływ jednego elementu na interpretację innego nie musi być stały, lecz może zależeć od aktualnego kontekstu całej sekwencji.

Podobieństwo do Hofstadterowskiej płynności jest tutaj funkcjonalne, a nie architektoniczne. W Slipnecie kontekst może zmieniać dostępność pojęć i relacji; w transformerze reprezentacja tokenu również zależy od otoczenia, ponieważ kolejne warstwy przekształcają jego stan w świetle informacji pochodzących z innych pozycji. Nie oznacza to, że mechanizm attention jest „poślizgiem pojęciowym” w sensie Copycata. Oznacza natomiast, że oba podejścia odrzucają uproszczony model znaku posiadającego zawsze tę samą, całkowicie niezależną reprezentację. Znaczenie operacyjne elementu zostaje współokreślone przez otoczenie.

Dobrze ilustruje to wieloznaczność języka naturalnego. Ten sam ciąg znaków może reprezentować inne znaczenie zależnie od zdania, dyskursu, domeny i intencji rozmówcy. Człowiek nie przetwarza więc słowa jako niezmiennego pojemnika semantycznego. Podobnie współczesny model językowy

konstruuje reprezentacje kontekstowe, w których ten sam token może zajmować różne regiony przestrzeni aktywacji w zależności od użycia.

Jest to istotny krok poza naiwne utożsamienie znaczenia z etykietą symbolu. Jednocześnie byłoby błędem kategorii, gdyby z kontekstowości reprezentacji automatycznie wywodzić istnienie pojęć psychologicznych identycznych z ludzkimi. Wielkie modele językowe uczą się bowiem nie jawnych definicji świata, lecz statystycznych regularności pozwalających przewidywać i generować kolejne elementy danych. W najprostszym ujęciu ich uczenie podstawowe jest podporządkowane zadaniu przewidywania brakujących lub kolejnych jednostek tekstu. Jednak realizacja tego celu na dostatecznie bogatym korpusie może wymagać powstawania reprezentacji, które kodują własności wykraczające daleko poza powierzchnię znaków. Aby przewidzieć odpowiednie słowo w tekście o chemii, historii, programowaniu czy relacjach społecznych, model musi – przynajmniej funkcjonalnie – wychwytywać regularności dotyczące struktur tych domen.

Nie wynika z tego, że buduje świat dokładnie tak jak człowiek, lecz że zadanie pozornie lokalne może premiować powstawanie struktur reprezentacyjnych o wysokim poziomie abstrakcji. Badania nad interpretowalnością dostarczają coraz mocniejszych dowodów na to, że wewnętrzne reprezentacje dużych modeli nie ograniczają się do powierzchownych cech słów. W 2024 roku badacze z Anthropic opisali metodę wydobywania z aktywacji modelu milionów cech odpowiadających zarówno konkretnym obiektom, jak i treściom abstrakcyjnym; część z nich była aktywna wielojęzycznie i multimodalnie. Jednocześnie autorzy wyraźnie zaznaczali, że zidentyfikowanie cechy nie oznacza pełnego zrozumienia obwodu, który ją wykorzystuje, ani całej mechaniki zachowania modelu. To zastrzeżenie jest niezwykle ważne w porównaniu z FARG.

Od projektowanej przejrzystości do wyłaniającej się złożoności

W modelach Hofstadtera reprezentacje są w dużej mierze projektowo interpretowalne; w LLM wyłaniają się one z treningu, a próby ich rekonstrukcji następują dopiero później i mają charakter eksperymentalny. Współczesne wyniki podważają jednak prosty zarzut, jakoby sieci neuronowe kodowały jedynie powierzchowne współwystępowanie słów. Badania nad interpretowalnością ujawniają cechy reagujące na abstrakcyjne konfiguracje znaczenia, a nie wyłącznie na konkretne leksemy. Jednocześnie wnioski te wynikają z analizy konkretnych rodzin modeli i metod pomiaru, więc nie można ich automatycznie rozszerzać na wszystkie LLM. Szczególnie ryzykowne byłoby zastąpienie terminu "feature" przez metafizycznie mocniejsze pojęcie "pojęcia w umyśle". Interpretowalna cecha to struktura aktywacji powiązana z określonym zakresem bodźców; pytanie,

czy i w jaki sposób stanowi ona pojęcie, wymaga osobnej teorii. To właśnie tutaj pytania Hofstadtera okazują się bardziej użyteczne niż jego konkretne rozwiązania architektoniczne. FARG badało, czy pojęcia są sztywne, jak kontekst zmienia ich relacje, w jaki sposób system odkrywa podobieństwo wyższego rzędu i jak konkurencyjne interpretacje prowadzą do reorganizacji reprezentacji. W nowoczesnym modelu językowym podobne problemy można analizować bez oczekiwania, że znajdziemy literalny odpowiednik Slipnetu.

Interesujące staje się raczej to, czy aktywacje reprezentujące określone własności systematycznie zmieniają się wraz z kontekstem, czy model potrafi stosować strukturę do nowych egzemplarzy oraz czy zmiana wewnętrznej reprezentacji ma przyczynowy wpływ na zachowanie. Mechanistyczna interpretowalność stanowi zatem współczesną odpowiedź na dawny postulat Hofstadtera, by zagłądać do procesu, a nie tylko oceniać produkt. W 2025 roku badacze z Anthropic opisali techniki śledzenia części ścieżek obliczeniowych prowadzących do konkretnych wyników modeli, podkreślając, że strategie wyuczone podczas treningu pozostają w znacznej mierze niejawne nawet dla twórców systemu. Sama możliwość mapowania pewnych obwodów nie oznacza jeszcze posiadania ogólnej teorii działania LLM. Świadczy jednak o zmianie metodologicznej: coraz mniej satysfakcjonujące jest pytanie "czy model odpowiedział?", a coraz ważniejsze – "co wewnątrz modelu przyczyniło się do tej odpowiedzi?".

Jest to niemal dokładne odwrócenie sytuacji z FARG. W Copycat zakres zachowań był bardzo wąski, natomiast mechanizm względnie przejrzysty. W modelu językowym zakres zachowań jest ogromny, ale przyczynowa interpretacja jego kompetencji pozostaje fragmentaryczna. Doszło więc do pewnego rodzaju wymiany epistemicznej: wzrost mocy behawioralnej odbył się kosztem utraty przejrzystości. Powoduje to, że współczesna AI jest jednocześnie znacznie bardziej imponująca technologicznie i trudniejsza do wykorzystania jako czysty model naukowy ludzkiego poznania.

Różnica ta dotyczy również genezy kategorii. Slipnet zawiera strukturę pojęciową przygotowaną przez człowieka. Model językowy nie otrzymuje natomiast jawnej instrukcji, że "pies", "kot" i "koń" tworzą jedną klasę semantyczną, że ironia różni się od literalności lub że rozwiązanie matematyczne może być strukturalnie analogiczne do problemu spoza matematyki. Regularności te muszą zostać w dużej mierze wyprowadzone z danych i procesu optymalizacji. Oznacza to niewątpliwie większy stopień autonomii reprezentacyjnej w jednym sensie: człowiek nie wpisuje ręcznie wszystkich pojęć. Nie oznacza jednak pełnej autonomii epistemicznej.

Rozproszone sprawstwo i zmienność kompetencji analogicznych

Człowiek wciąż decyduje o architekturze, danych, procedurach uczenia, celach treningowych i metodach post-treningu, a same dane niosą w sobie ogromny, skumulowany dorobek poznawczy ludzkości. Właśnie dlatego klasyczny problem **credit assignment** powraca w nowej skali. Gdy model rozwiązuje skomplikowaną analogię, trudno jednoznacznie wskazać źródło tej zasługi. Wynik zależy bowiem od architektury transformera, zbiorów treningowych, optymalizacji, post-treningu, aktualnego promptu oraz samych obliczeń wykonywanych podczas inferencji.

Nie byłoby uczciwe twierdzić, że programista po prostu „wpisał odpowiedź”, gdyż w zdecydowanej większości przypadków nie stało się tak w klasycznym sensie. Równie nieuzasadnione byłoby jednak pominięcie zależności zdolności modelu od ludzkiej kultury utrwalonej w danych. Sprawstwo poznawcze staje się układem przyczyn rozproszonych. To z kolei komplikuje dawne rozróżnienie pomiędzy kombinatoryką a kreatywnością.

Sieć uczy się na istniejących danych, lecz rezultat nie musi być prostym sklejeniem pojedynczych fragmentów treningowych. Badania nad wpływem danych wskazują, że w przypadku typowych odpowiedzi wpływ ten bywa rozproszony po wielu przykładach, zamiast sprowadzać się do literalnego odtworzenia jednej sekwencji. Nie dowodzi to jeszcze kreatywności w sensie silnym, podważa jednak naiwny obraz modelu jako katalogu gotowych zdań odnajdującego właściwy zapis. Jeszcze ciekawszy staje się problem analogii. W 2023 roku Webb, Holyoak i Lu wykazali, że GPT-3 osiągał w wielu zadaniach analogicznych wyniki zbliżone do ludzkich lub je przewyższające – dotyczyło to również problemów skonstruowanych tak, by testować abstrakcyjne relacje, a nie jedynie odtwarzanie typowych fraz.

Było to empirycznie istotne, ponieważ zdolność analogiczna zajmuje centralne miejsce w teoriach Hofstadtera. System powstały poza tradycją FARG zaczął wykazywać kompetencję, którą FARG uznaje za jeden z rdzeni inteligencji. Nie można jednak na tej podstawie ogłosić rozwiązania problemu analogii. Reanaliza opublikowana w 2025 roku wskazała na dużą zmienność wyników modeli i podważyła zbyt śmiało uogólnienia wyprowadzane z najlepszych rezultatów. To napięcie jest metodologicznie bardzo produktywne. LLM nie są ani systemami „pozbawionymi analogii”, ani stabilnymi maszynami analogicznymi, równoważnymi człowiekowi we wszystkich warunkach. Ich kompetencje są nierówne i zależne od zadania, formatu, promptu, modelu oraz warunków ewaluacji.

Właśnie ta zmienność może być jednym z najciekawszych punktów spotkania z filozofią FARG. Hofstadter podkreślał, że inteligencja ujawnia się nie tylko w poprawnej odpowiedzi, lecz również w profilu błędów i sposobie reagowania na zmianę kontekstu. Jeśli model rozwiązuje analogię w jednym sformułowaniu, ale zawodzi po niewielkiej zmianie powierzchniowej, może to sugerować, że jego reprezentacja pozostaje częściowo przywiązana do formy zadania. Jeżeli natomiast zachowuje strukturę rozwiązania pomimo radykalnej zmiany słownictwa i domeny, argument za reprezentacją relacyjną staje się silniejszy. Nie istnieje więc pojedynczy wynik, który rozstrzygałby kwestię „posiadania analogii” – potrzebny jest profil generalizacji.

Podobny problem dotyczy **in-context learning**, jednej z najbardziej charakterystycznych zdolności współczesnych modeli. Model otrzymuje w kontekście kilka przykładów nowej reguły lub sposobu zadania i często potrafi dostosować do nich dalsze zachowanie bez trwałej aktualizacji parametrów. Funkcjonalnie przypomina to lokalną rekonfigurację systemu pod wpływem bieżącego środowiska, choć oczywiście nie jest to odpowiednik Slipnetu.

Plastyczność stanu a funkcjonalna zbieżność z Workspace

Jest to zjawisko szczególnie interesujące z perspektywy płynnych pojęć, gdyż pokazuje, że stałe parametry modelu mogą generować dynamiczne zachowania silnie zależne od chwilowego kontekstu. Należy tu rozróżnić plastyczność parametrów od plastyczności stanu. Podczas treningu zmieniają się parametry modelu; podczas zwykłej rozmowy mogą one pozostawać zamrożone, lecz zmieniają się stany aktywacji i zawartość kontekstu, co wystarcza do wywołania znacznych zmian w zachowaniu. Inteligencja kontekstowa nie wymaga więc permanentnego uczenia po każdym nowym zdaniu – może być realizowana przez dynamiczne konstruowanie reprezentacji w obrębie aktualnego przebiegu. Rozróżnienie to jest istotne również w odniesieniu do człowieka, który nie przebudowuje całej pamięci długotrwałej przy każdej lokalnej hipotezie, lecz utrzymuje tymczasowe struktury robocze.

W tym miejscu można dostrzec odległe podobieństwo do Workspace FARG. Workspace ten służył budowaniu tymczasowej organizacji problemu na bazie trwalszych zasobów konceptualnych. Analogia ta wymaga jednak ostrożności. W transformerze kontekst jest przetwarzany przez warstwy sieci i przechowywany w rozproszonych aktywacjach; w Copycat Workspace stanowi jawnie zaprojektowaną strukturę zawierającą grupy, wiązania i mosty analogiczne. Podobna jest zatem funkcja – utrzymanie roboczej reprezentacji – a nie mechanizm.

Jeszcze ciekawsze są wyniki najnowszych badań nad interpretowalnością. W lipcu 2026 roku zespół Anthropic opisał w modelach Claude zjawisko określone jako „global workspace”: niewielki zbiór wewnętrznych reprezentacji miał pełnić szczególną rolę w udostępnianiu treści innym procesom modelu. Autorzy wyraźnie odnosili to do funkcjonalnego pojęcia globalnej dostępności, a nie do rozstrzygnięcia problemu świadomości fenomenalnej. Z punktu widzenia analizy autora książki wynik ten jest interesujący właśnie dlatego, że pokazuje możliwość spontanicznego wyłaniania się wewnętrznych struktur pełniących rolę nieprzewidzianą w prostym opisie celu treningowego. Nie wolno jednak twierdzić, że LLM posiada Workspace Hofstadtera lub że wynik ten dowodzi świadomości. Byłoby to przekroczenie danych.

Można jedynie zauważyć, że współczesne badania identyfikują w wielkich sieciach funkcjonalnie wyróżnione reprezentacje, które integrują i udostępniają informacje w sposób przypominający niektóre wcześniejsze idee architektury poznawczej. Podobieństwo to może być teoretycznie inspirujące, lecz różnice mechanistyczne pozostają zasadnicze. Dalej idą badania próbujące tłumaczyć stany aktywacji na język naturalny. W maju 2026 roku Anthropic opisał Natural Language Autoencoders – narzędzie mające tworzyć językowe opisy informacji zawartych w aktywacjach modeli. Autorzy przedstawiali przykłady wcześniejszego planowania rymów czy wewnętrznych stanów związanych z zachowaniem modelu, analizując jednocześnie ograniczenia tej metody. Tego typu narzędzia są interesujące, ponieważ zmniejszają przepaść pomiędzy obserwowalnym tekstem a ukrytą przestrzenią aktywacji. Nadal jednak nie są „oknem do umysłu” w prostym sensie.

Funkcjonalne pojęcia zamiast binarnego rozumienia

Każdy dekodery aktywacji jest kolejnym modelem i wymaga walidacji. Sama obecność reprezentacji wewnętrznych prowadzi do zasadniczego pytania: czy można uznać, że LLM posiada pojęcia? Odpowiedź zależy od przyjętej definicji. Jeśli za pojęcie uznamy stabilną strukturę reprezentacyjną, która umożliwia generalizację między różnymi wystąpieniami i wpływa na zachowanie, istnieją przesłanki, by mówić o funkcjonalnych odpowiednikach właściwości pojęciowych. Jeśli jednak wymagamy ucieleśnionego zakorzenienia, historii interakcji ze światem, świadomego doświadczenia lub udziału w społecznych praktykach nadawania znaczeń, sam model tekstowy może nie spełniać tych kryteriów. Spór ten nie daje się rozstrzygnąć jedną etykietą.

Podobnie wieloznaczne jest słowo „rozumienie”. W jednej tradycji oznacza ono zdolność poprawnego stosowania struktury do nowych przypadków; w innej wymaga możliwości wyjaśnienia przyczyn i konsekwencji, a jeszcze w kolejnej wiąże się z intencjonalnością lub

doświadczeniem świata. Model językowy może spełniać część tych kryteriów, pomijając inne. Z naukowego punktu widzenia bardziej produktywnie jest zatem rozłożenie rozumienia na operacjonalizowane zdolności niż prowadzenie sporu o binarny status systemu.

Hofstadterowska filozofia okazuje się tutaj paradoksalnie sprzymierzeńcem ostrożności, gdyż cały program FARG wyrósł z nieufności wobec powierzchownych kryteriów inteligencji. Płynne zachowanie językowe może być istotnym dowodem kompetencji, lecz nie powinno automatycznie zamykać pytania o strukturę wewnętrzną. Jednocześnie współczesne LLM stanowią silny kontrargument dla zbyt prostego poglądu, jakoby system uczony statystycznie na tekście z definicji nie mógł osiągnąć wysokopoziomowych reprezentacji. Empiria pokazała, że "statystyczny" nie musi znaczyć "powierzchny". To jedna z najważniejszych korekt, jakie współczesność wnosi do sporów z końca XX wieku.

Statystyczne uczenie i abstrakcyjna reprezentacja nie są przeciwieństwami; statystyczny proces optymalizacji może doprowadzić do powstania struktur funkcjonalnie abstrakcyjnych. Problem nie brzmi więc: „reguły czy statystyka?”, lecz: „jakiego rodzaju struktury wyłaniają się ze statystycznego uczenia i w jakich warunkach pozostają zdolne do generalizacji?”. Ta zmiana perspektywy jest porównywalna z przejściem od dyskusji o pojedynczym symbolu do analizy organizacji całej przestrzeni reprezentacyjnej. W tym miejscu użyteczne okazuje się pojęcie superpozycji, rozwijane we współczesnej interpretowalności. Jednostka sieci neuronowej nie musi odpowiadać pojedynczemu pojęciu – wiele cech może być kodowanych w nakładających się wzorcach aktywacji. Badania dictionary learning próbują rozdzielać te wzorce na bardziej interpretowalne komponenty. To radykalnie inny sposób reprezentowania wiedzy niż klasyczna baza symboliczna, choć paradoksalnie odpowiada jednej z intuicji Hofstadtera: znaczenie wysokiego poziomu nie musi być umiejscowione w jednym centralnym symbolu lub module.

Kodlet posiada zaprojektowaną rolę funkcjonalną; neuron czy feature w LLM nie jest agentem realizującym samodzielny mini-program w sensie FARG. Nie należy antropomorfizować lokalnych aktywacji. Możemy mówić o rozproszonej przyczynowości, konkurencji sygnałów i integracji informacji, ale terminologia agentowa może w tym miejscu zaciemniać zamiast wyjaśniać. Równie ostrożnie trzeba traktować attention. Nazwa sugeruje podobieństwo do ludzkiej uwagi, lecz techniczny mechanizm nie jest cyfrową wersją psychologicznego kierowania uwagą – to matematyczna operacja dynamicznego ważenia relacji pomiędzy reprezentacjami. Może ona pełnić funkcje częściowo analogiczne do selekcji informacji, jednak fenomenologiczna i neuronalna ludzka uwaga jest zjawiskiem znacznie bogatszym. Historia AI pokazuje, jak łatwo termin techniczny zaczyna sugerować nieuprawnioną tożsamość psychologiczną.

To zastrzeżenie nie pomniejsza znaczenia transformera; wręcz przeciwnie – pozwala lepiej dostrzec jego niezwykłość. System został zbudowany bez ręcznego wpisywania ogromnej liczby pojęć, a mimo to wyuczył struktury pozwalające na wykonywanie zadań, których projektanci nie implementowali oddzielnie. To właśnie tutaj pojawia się prawdziwy problem emergencji. Rezultat nie jest jawnie zakodowaną regułą w klasycznym sensie, ale również nie powstaje „z niczego”. Jest konsekwencją architektury, danych, optymalizacji i skali.

Generalizacja strukturalna jako miara rzeczywistej emergencji

Współczesna dyskusja o emergent capabilities pokazuje, jak ostrożnie należy posługiwać się samym terminem emergencji. Nagły skok w wynikach konkretnego benchmarku może być jedynie efektem sposobu progowania danych, podczas gdy rzeczywista wydajność (underlying performance) rośnie w sposób ciągły. W innych przypadkach rzeczywiście mamy do czynienia z jakościową zmianą dostępnych strategii. Nie wolno zatem traktować każdego zaskoczenia przy skalowaniu jako dowodu pojawienia się nowej ontologicznej własności systemu. Hofstadterowska metafora emergencji również wymaga takiej dyscypliny: odnosi się ona do relacji między poziomami organizacji, a nie o magiczne przekroczenie praw obliczenia. Znacznie bardziej interesujące od pytania o to, czy dana zdolność „pojawiła się nagle”, jest zagadnienie, czy model nabył reprezentację pozwalającą przenosić kompetencję poza dane bezpośrednio podobne do treningowych. Kluczowym testem staje się tutaj generalizacja strukturalna. Jeśli model potrafi zastosować tę samą relację w nowej domenie o radykalnie innej powierzchni, otrzymujemy silniejszy dowód na zdolność abstrakcji, niż gdy odpowiada poprawnie w zadaniu bliskim wzorcom treningowym.

Trudno jest jednak całkowicie wyeliminować problem skażenia danych. Wielkie modele uczone są na tak obszernych korpusach, że badacz często nie może z absolutną pewnością stwierdzić, czy dany problem nie wystąpił już wcześniej w zbiorze treningowym. Dlatego ocena inteligencji LLM wymaga coraz staranniejszego projektowania zadań: nowych, proceduralnie generowanych, kontrfaktycznych lub opartych na sztucznych regułach, których model nie mógł po prostu zapamiętać.

Synteza FARG i LLM jako program badawczy

W tym ujęciu mikrodomeny FARG odzyskują współczesną wartość metodologiczną. Mały świat może służyć jako kontrolowane laboratorium generalizacji, o ile jego strukturę można systematycznie zmieniać. Jednocześnie dzisiejsze systemy obnażają ograniczenia mikrodomen Hofstadtera: płynność działająca w starannie zaprojektowanym mikroświecie nie jest jeszcze płynnością w otwartym języku i kulturze. Modele językowe udowodniły technologicznie, że szerokość domeny ma znaczenie – kompetencje mogą wyłaniać się z kontaktu z ogromną różnorodnością tekstów i zadań, a nie tylko z perfekcyjnie przejrzystej architektury w wąskim środowisku. FARG wygrywa zatem pod względem interpretowalności, natomiast LLM pod kątem zasięgu generalizacji behawioralnej.

To napięcie może prowadzić do nowej syntezy. Zamiast wybierać między mikrodomeną a skalą, można potraktować współczesne modele jako potężne generatory hipotez o inteligencji, które następnie należy badać metodami zbliżonymi do ambicji FARG: poprzez tworzenie kontrolowanych doświadczeń, manipulowanie reprezentacjami, obserwację skutków przyczynowych i analizę struktury błędów. Wówczas technologia nie zastępuje kognitywistyki, lecz staje się jej nowym przedmiotem eksperymentalnym. Kluczowym obszarem takich badań jest rozróżnienie pomiędzy reprezentacją latentną a jawnym rozumowaniem językowym. Model może bowiem posiadać informacje zakodowane w aktywacjach, których nie ujawnia bezpośrednio w odpowiedzi; współczesne techniki interpretowalności próbują badać właśnie tę warstwę.

Wyniki dotyczące planowania przed wygenerowaniem tekstu czy istnienia stanów dostępnych wewnątrznie – niewidocznych w bieżącym outputcie – komplikują naiwny obraz LLM jako systemu, który „myśli wyłącznie słowami” i istnieje tylko na poziomie kolejno produkowanych tokenów. Nie oznacza to jednak, że powinniśmy bezkrytycznie nazywać te stany „myśłami”. Termin ten jest metaforycznie wygodny, lecz w ujęciu naukowym należy mówić przede wszystkim o aktywacjach i reprezentacjach pełniących określone funkcje. Taka ostrożność pozwala uniknąć zarówno redukcjonizmu, jak i antropomorfizacji. Nie trzeba twierdzić, że system posiada ludzką świadomość, aby uznać, iż wewnętrzna organizacja jego obliczeń jest znacznie bogatsza niż prosty łańcuch słów.

Z perspektywy FARG szczególnie intrygująca jest możliwość występowania w tych reprezentacjach czegoś funkcjonalnie podobnego do konkurujących presji pojęciowych. Wygenerowanie odpowiedzi jest wynikiem współoddziaływania licznych sygnałów zakodowanych w parametrach i aktywacjach. Interpretowalność zaczyna ujawniać, że zachowanie może zależeć od różnych cech o przeciwnych tendencjach. Nie jest to Coderack, lecz ponownie pojawia się zasada rozproszonej kontroli: rezultat

nie musi być produktem jednego centralnego symbolicznego modułu zawierającego pełną odpowiedź. Równocześnie LLM radykalnie komplikują Hofstadterowski ideał mechanizmu zrozumiałego. Jeśli inteligencja może wyłonić się z miliardów parametrów, możliwe staje się stworzenie systemu, który zachowuje się inteligentnie szybciej, niż potrafimy sformułować pełną teorię jego działania.

To sytuacja epistemologicznie niezwykle. W klasycznym programowaniu rozumienie mechanizmu poprzedza jego wykonanie. W uczeniu maszynowym możemy zaprojektować procedurę uczenia, której końcowego produktu nie potrafimy szczegółowo przewidzieć ani wyjaśnić. Technologia spełniła zatem Hofstadterowską ideę emergencji bardziej dosłownie, niż zapewne przewidywali twórcy FARG, lecz zrobiła to kosztem drugiego ideału – przejrzystości mechanizmu. Systemy współczesne pokazują, że można uzyskać zdumiewającą kompetencję bez ręcznego zaprojektowania każdej reprezentacji. Nie dowodzą jednak jeszcze, że rozumiemy, w jaki sposób ta kompetencja powstaje na poziomie poszczególnych mechanizmów. Dlatego kluczowe pytanie nie może dziś brzmieć: „czy LLM są tylko statystyką?”. Wszystkie systemy uczenia maszynowego wykorzystują mechanizmy statystyczne, podobnie jak ludzki mózg funkcjonuje w świecie procesów biologicznych i probabilistycznych.

Słowo „statystyka” nie rozstrzyga poziomu abstrakcji powstającej reprezentacji. Pytanie właściwe brzmi: jakiego rodzaju struktury statystyczne są reprezentowane, jak stabilnie można je przenosić i czy system potrafi reorganizować je pod wpływem nowych warunków?

Tu powraca Hofstadterowskie pojęcie płynności. Być może płynne pojęcie nie musi być zapisane w jawnej sieci węzłów i poślizgów, a funkcjonalna płynność może wynikać z dynamicznej geometrii rozproszonych reprezentacji, zmieniających swój wpływ wraz z kontekstem. Aby jednak taką tezę naukowo obronić, nie wystarczy wskazać na efektowną rozmowę z modelem. Trzeba wykazać odporność na perturbacje, transfer między domenami, zachowanie relacji przy zmianie powierzchni oraz przyczynową rolę reprezentacji wewnętrznych. Właśnie dlatego model językowy jest dla teorii Hofstadtera zarazem wyzwaniem i potwierdzeniem jej najgłębszej intuicji. Jest wyzwaniem, gdyż pokazuje, że konkretna architektura FARG nie była konieczna do osiągnięcia znacznej części kompetencji analogicznych i generatywnych. Jest potwierdzeniem w słabszym, lecz ważniejszym sensie, ponieważ współczesna AI ponownie zmusza nas do pytania o reprezentację, kontekst, analogię, selekcję i emergencję, a nie tylko o szybkość obliczeń. Najbardziej naukowo interesujący wniosek nie brzmi zatem ani „Hofstadter miał rację”, ani „Hofstadter się mylił”.

Jego teoria została skonfrontowana z klasą systemów, których rozwój zmienił samą przestrzeń możliwych odpowiedzi. Okazało się, że reprezentacje mogą być zarazem bardziej rozproszone, mniej jawne i znacznie bardziej funkcjonalnie elastyczne, niż sugerowały spory między

symbolizmem a konekcjonizmem. Jednocześnie nierówność wyników analogicznych i wciąż ograniczona przejrzystość mechanistyczna pokazują, że behawioralne podobieństwo do człowieka nadal nie rozstrzyga pytania o organizację poznawczą.

Inteligencja poszarpana jako miara stabilności generalizacji

Stąd wynika następny etap analizy. Skoro modele językowe potrafią wykonywać liczne zadania analogiczne, należy przejść od pytania **czy** do pytania **jak stabilnie i jak głęboko**. Konieczne staje się zbadanie granicy pomiędzy rozpoznawaniem relacji a dopasowaniem powierzchniowym, wpływu formatu i promptu, problemu transferu strukturalnego, kontrfaktyczności oraz charakterystycznej "poszarpanej" kompetencji systemów, które potrafią rozwiązać niezwykle trudny problem, a następnie potknąć się o jego pozornie łatwiejszą odmianę. To właśnie tam współczesne modele spotykają się najpełniej z Hofstadterowskim przekonaniem, że prawdziwa miara inteligencji nie tkwi w pojedynczym imponującym wyniku, lecz w strukturze generalizacji i błęd.

Poszarpana inteligencja: analogia, transfer strukturalny i granice generalizacji modeli językowych

Najtrudniejszym pytaniem dotyczącym rozumowania analogicznego modeli językowych nie jest już to, czy potrafią one rozwiązać zadanie analogiczne. Potrafią. Problem zaczyna się poziom wyżej: **jakiego rodzaju kompetencję ujawnia poprawne rozwiązanie, jak daleko kompetencja ta przenosi się poza konkretną formę zadania oraz jak zachowuje się wtedy, gdy powierzchniowe wskazówki zostają celowo oddzielone od struktury relacyjnej**. To właśnie tutaj filozofia płynnych pojęć Hofstadtera okazuje się nadal użyteczna.

Copycat nie był bowiem pomyślany jako maszyna produkująca jedną poprawną odpowiedź. Miał ujawniać, jak konkurencyjne interpretacje powstają, są wzmacniane, porzucane i rekonstruowane, a przede wszystkim – które własności sytuacji system uznaje za głębokie, a które pozwala sobie "przesunąć". Współczesne badania nad LLM stawiają właściwie tę samą kwestię, choć w innej aparaturze pojęciowej. Jeśli model rozwiązuje problem, ponieważ rozpoznał jego strukturę relacyjną, powinien zachować znaczną część kompetencji po zmianie nazw, kolejności elementów, narracji, dziedziny semantycznej albo powierzchniowego formatu. Jeżeli sukces zależy natomiast przede wszystkim od bliskości do wyuczonych wzorców językowych, niewielka perturbacja może spowodować gwałtowne załamanie wyniku.

W praktyce oba mechanizmy mogą współistnieć. Model może posiadać pewne reprezentacje abstrakcyjne, a równocześnie korzystać z heurystyk powierzchniowych; może generalizować w jednej rodzinie transformacji, a pozostawać kruchy w innej. Inteligencja nie musi więc być własnością jednorodną. Znany eksperyment Webba, Holyoaka i Lu był przełomowy właśnie dlatego, że zakwestionował prostą tezę, według której duży model językowy nie jest zdolny do prawdziwie abstrakcyjnego rozumowania analogicznego. GPT-3 badano między innymi w zadaniach opartych na strukturach macierzy Ravena, analogiach literowych, werbalnych i narracyjnych. W wielu warunkach model osiągał rezultaty porównywalne z ludźmi, a czasem je przewyższał. Autorzy interpretowali wyniki jako dowód pojawienia się istotnej zdolności do indukcji abstrakcyjnych wzorców i rozwiązywania problemów analogicznych bez bezpośredniego treningu na konkretnym zadaniu. Rezultat ten ma duże znaczenie dla programu niniejszego wywodu, ponieważ analogia zajmuje w teorii Hofstadtera miejsce centralne.

Jeśli system powstały przez predykcyjne uczenie języka, bez jawnego Slipnetu i Coderacka, zaczyna zachowywać się jak podmiot zdolny do odnajdywania analogii, oznacza to co najmniej, że funkcjonalna kompetencja analogiczna może zostać osiągnięta inną drogą architektoniczną. Nie rozstrzyga to jeszcze, czy mechanizm modelu i mechanizm ludzkiej analogii są podobne. Zmusza jednak do porzucenia tezy, że bez jawnej manipulacji relacjami analogia jest z definicji niedostępna. Jednocześnie późniejsza analiza wyników tych samych modeli wprowadziła istotne zastrzeżenie. W 2025 roku de Varda, Saponaro i Marelli zwrócili uwagę na dużą zmienność rezultatów LLM w zadaniach analogicznych i zakwestionowali zbyt mocne uogólnienie, według którego wysoki wynik w wybranych warunkach pozwala przypisać modelowi stabilną kompetencję analogiczną porównywalną z ludzką. Warto potraktować ten spór nie jako prostą korektę wcześniejszego "entuzjazmu", lecz jako ilustrację problemu badań nad inteligencją sztuczną: **zachowanie modelu może być bardzo wysokie w jednym wycinku przestrzeni zadań i jednocześnie niestabilne po przesunięciu warunków pomiaru**. Właśnie tę właściwość współczesnych systemów dobrze oddaje termin "jagged intelligence" – inteligencja poszarpana.

Stanford AI Index 2026 zwraca uwagę na wyraźną niespójność profilu kompetencji najnowszych systemów: modele potrafią osiągać wyniki odpowiadające najwyższemu poziomowi zawodów matematycznych, a jednocześnie pozostawać wyraźnie słabsze od ludzi w zadaniach percepcyjnych, które z ludzkiego punktu widzenia wydają się znacznie prostsze, jak niezawodne odczytywanie analogowego zegara.

Jagged frontier podważa liniową skalę inteligencji

Raport interpretuje tę nierówność jako przykład jagged frontier – granicy kompetencji, która nie przesuwa się równomiernie wraz ze wzrostem ogólnej wydajności modeli. Obserwacja ta jest teoretycznie istotniejsza niż sama lista benchmarków. Gdybyśmy wyobrażali sobie inteligencję jako jednowymiarową skalę, naturalne byłoby oczekiwanie, że system zdolny do rozwiązywania problemu „trudniejszego” będzie niezawodnie radził sobie z zadaniami „łatwiejszymi”.

Tymczasem współczesne modele podważają ten porządek. Trudność nie jest bowiem cechą absolutną danego zadania, lecz wynika z relacji między jego strukturą a architekturą systemu. Problem banalny dla człowieka może wymagać rodzaju reprezentacji, której model nie posiada w sposób dostatecznie stabilny; z kolei zagadnienie bardzo trudne dla ludzi może odpowiadać strukturze, którą ogromny model potrafi sprawnie wydobyć ze swojego treningu. Zjawisko to nie jest całkowicie obce psychologii człowieka – ludzkie kompetencje również są nierówne. Ktoś może być wybitnym matematykiem, a jednocześnie słabo orientować się w przestrzeni; znać wiele języków, lecz nie posiadać zdolności muzycznych, lub dysponować ogromną wiedzą deklaratywną, podejmując błędne decyzje pod presją emocjonalną. Różnica tkwi jednak w skali i nieintuicyjności pewnych dysproporcji AI. W systemach sztucznych spotykamy kombinacje sprawności i bezradności, które nie odpowiadają naturalnej hierarchii ludzkiego rozwoju poznawczego. To właśnie dlatego antropomorficzne pojęcie „poziomu inteligencji” bywa mylące.

FARG dostarcza narzędzi pozwalających zrozumieć to zjawisko. W jego modelach wynik zależy od dostępnych reprezentacji, presji pojęciowych i sposobu reorganizacji lokalnych operacji. System nie jest „inteligentny” w oderwaniu od mikrodomeny; jego płynność istnieje względem określonej przestrzeni reprezentacyjnej. Autor książki podkreśla, że mikrodomeny są zarówno metodologicznym atutem, jak i ograniczeniem FARG – pozwalają badać mechanizm w szczegółach, ale nie dowodzą automatycznie, że ten sam mechanizm przeskaluje się do otwartego świata. Współczesne LLM rozwiązują problem skali znacznie skuteczniej, lecz odziedziczyły inny rodzaj ograniczenia: ich kompetencje są niezwykle szerokie, ale nie zawsze układają się w jednolity i przewidywalny profil. Sugeruje to, że „ogólność” inteligencji należy rozumieć ostrożniej. System może być ogólny pod względem liczby domen, do których można go zastosować, a jednocześnie nie być ogólny w sensie posiadania jednego stabilnego mechanizmu rozumowania, działającego identycznie niezależnie od powierzchni zadania. Model potrafi transferować strukturę pomiędzy bardzo odległymi przypadkami, lecz niekiedy pozostaje wrażliwy na pozornie kosmetyczne zmiany prezentacji. Ogólność zakresu nie gwarantuje zatem ogólności mechanizmu. Kluczowe staje się tutaj rozróżnienie między generalizacją interpolacyjną a generalizacją strukturalną.

Stabilność strukturalna jako miara inteligencji

W pierwszym przypadku nowa sytuacja znajduje się w pobliżu rozkładu znanych przykładów: model rozpoznaje kombinację właściwości zbliżoną do tej, z którą zetknął się wcześniej. W drugim musi wyodrębnić bardziej abstrakcyjną zasadę i przenieść ją na grunt przypadku różniącego się powierzchniowo. Granica między tymi kategoriami nie jest ostra; ogromny zbiór treningowy może sprawić, że to, co badacz uznaje za „daleki transfer”, w przestrzeni reprezentacyjnej modelu znajduje się bliżej danych treningowych, niż przypuszczamy. To jeden z powodów, dla których interpretacja benchmarków LLM jest tak trudna.

W ludzkiej analogii również obserwujemy silny wpływ podobieństwa powierzchniowego. Klasyczne badania Gentner, Holyoaka i innych badaczy pokazywały, że ludzie często nie przenoszą znanego rozwiązania na nowy problem, jeśli oba przypadki różnią się oprawą semantyczną, mimo że ich struktura relacyjna pozostaje analogiczna. Dopiero dodatkowa wskazówka skłania ich do zauważenia wspólnej relacji. Nie należy więc wymagać od maszyny absolutnej niewrażliwości na powierzchnię, aby uznać ją za zdolną do analogii. Istotna jest raczej charakterystyka tego, jak powierzchnia konkuruje ze strukturą. Hofstadter ujmował ten problem poprzez dynamikę głębokości pojęciowej: płytkie cechy powinny być łatwiej porzucane, natomiast głębokie relacje – bardziej odporne na poślizg.

Współczesny odpowiednik takiego testu można skonstruować poprzez systematyczne perturbowanie problemu. Jeśli zmienimy nazwy obiektów, język, kolejność prezentacji, styl narracji i nieistotne szczegóły, a model zachowa rozwiązanie, rośnie prawdopodobieństwo, że korzysta on z reprezentacji strukturalnej. Jeśli natomiast zmiana jednego powierzchniowego sygnału radykalnie zmienia odpowiedź, ujawnia się zależność od wskazówki, która z perspektywy człowieka mogła wydawać się drugorzędna. Należy jednak zachować ostrożność również w odwrotnym kierunku.

Odporność behawioralna na perturbacje nie dowodzi sama w sobie, że model skonstruował identyczną reprezentację jak człowiek. Różne mechanizmy mogą prowadzić do podobnych wzorców odpowiedzi; to podstawowy problem niedookreślenia teorii przez zachowanie. Jeśli dwa systemy rozwiązują tę samą klasę zadań, potrzebujemy dodatkowych eksperymentów przyczynowych, aby sprawdzić, jakie reprezentacje rzeczywiście odpowiadają za wynik. Z tego względu przyszłość badań nad analogią w AI leży nie tyle w powiększaniu liczby pojedynczych pytań testowych, co w projektowaniu rodzin transformacji. Zamiast sto razy pytać o rozwiązanie konkretnej analogii, można stworzyć jeden problem strukturalny i systematycznie modyfikować wszystko, co nie powinno mieć znaczenia: nazwy, język, modalność, perspektywę, kolejność danych, narrację czy dystans semantyczny między domenami. Obserwując, które transformacje pozostawiają odpowiedź

stabilną, a które powodują jej załamanie, otrzymujemy nie pojedynczy wynik, lecz mapę inwariancji systemu. Podejście to jest niezwykle bliskie filozofii Copycata.

Prawdziwym pytaniem nie jest zatem „czy odpowiedź brzmi X?”, lecz „jakie transformacje system uważa za zachowujące istotę problemu?”. Autor książki traktuje rozwiązanie Copycata jako rezultat negocjowania między możliwymi interpretacjami, a nie zastosowanie raz zdefiniowanej reguły. Współczesny LLM można badać w podobny sposób, choć jego wewnętrzny proces jest inny: interesuje nas nie tylko sama odpowiedź, ale przede wszystkim to, jak zmiana reprezentacji wejściowej wpływa na strukturę rozwiązania. Szczególnie wartościowym testem pozostaje tutaj transfer między domenami.

Pośrednie ugruntowanie a poszarpany profil kompetencji

Jeżeli model rozpoznaje pewien schemat przyczynowy w historii wojskowej i potrafi zastosować go do problemu medycznego albo ekonomicznego, argument za abstrakcyjną reprezentacją relacji zyskuje na sile. Nie musi to oznaczać identyczności ludzkiego i maszynowego „rozumienia”, lecz pokazuje, że zachowanie nie daje się wyjaśnić wyłącznie lokalnym podobieństwem słów. Analogicznie człowiek dostrzega pełną wartość analogii dopiero wtedy, gdy relacja może zostać odseparowana od treści źródłowej. Tutaj jednak pojawia się istotna asymetria między człowiekiem a modelem językowym. Człowiek dysponuje stosunkowo ograniczonym doświadczeniem tekstowym, ale jego pojęcia są osadzone w percepcji, działaniu, cielesności i długotrwałej historii interakcji ze światem. LLM posiada niewyobrażalnie większą ekspozycję na tekst, lecz niekoniecznie ten sam rodzaj ugruntowania. Badanie opublikowane w **Nature Human Behaviour** w 2025 roku wskazało interesującą granicę: nieugruntowane modele językowe potrafiły odtwarzać wiele niesensomotorycznych własności ludzkich pojęć, jednak znacznie gorzej radziły sobie z odzyskiwaniem cech zależnych od doświadczenia sensoryczno-motorycznego.

Wynik ten jest istotny, ponieważ podważa zarówno naiwny tekstualizm, jak i jego prostą negację. Z jednej strony ogromna część struktury pojęciowej może być rekonstruowana na podstawie regularności języka, gdyż język jest produktem ludzi posiadających ciała i działających w świecie. Tekst zawiera więc pośredni ślad struktury doświadczenia. Z drugiej strony nie wszystkie własności pojęć są równie łatwo odzyskiwane w ten sposób.

Informacja sensoryczno-motoryczna może wymagać bezpośredniego ugruntowania albo przynajmniej bogatszego treningu multimodalnego. Prowadzi to do teorii **pośredniego**

ugruntowania. Model tekstowy nie doświadcza ognia, ale analizuje miliony zdań napisanych przez ludzi, którzy go widzieli, czuli jego ciepło, obawiali się poparzenia i wykorzystywali go w praktyce. Język jest więc skompresowanym depozytem doświadczeń podmiotów ucieleśnionych.

Model może nauczyć się znacznej części struktury pojęcia „ogień” bez osobistego kontaktu z płomieniem. Nie oznacza to jednak, że posiada wszystkie komponenty ludzkiej reprezentacji tego zjawiska. Jego pojęcie może być funkcjonalnie wystarczające dla jednych zadań i jakościowo uboższe dla innych. Ta asymetria pomaga zrozumieć poszarpany profil kompetencji. Model może wykazywać niezwykłą biegłość tam, gdzie struktura zadania została bogato zakodowana w kulturze językowej, a jednocześnie zawodzić w problemach zależnych od precyzyjnej geometrii, sensorimotoryki lub interakcji ze środowiskiem. Wraz z rozwojem systemów multimodalnych granice te mogą się przesuwąć, lecz ogólna zasada pozostanie niezmienna: **profil inteligencji odzwierciedla strukturę doświadczenia, do którego system miał dostęp podczas uczenia i działania.** Dlatego nie powinniśmy oczekiwać, że kolejne modele będą rozwijały się według ludzkiej kolejności ontogenetycznej. Dziecko najpierw poznaje świat poprzez działanie i percepcję, a dopiero później opanowuje język i formalne abstrakcje. LLM przeszły tę drogę odwrotnie: otrzymały ogromny świat języka, zanim uzyskały bogatszy kontakt z obrazem, dźwiękiem, przestrzenią czy działaniem. Ich hierarchia łatwych i trudnych problemów może zatem pozostawać nieintuicyjna z punktu widzenia psychologii rozwojowej. Właśnie dlatego określanie AI mianem „dziecka” albo „ucznia” jest często bardziej mylące niż wyjaśniające.

System może znać więcej faktów niż większość dorosłych, wykonywać zaawansowaną analizę kodu i jednocześnie nie posiadać elementarnych komponentów doświadczenia, które u człowieka pojawiają się przed nabyciem języka. To nie niedojrzały człowiek, lecz inna trajektoria konstrukcji kompetencji. Tak rozumiana poszarpana inteligencja ma również konsekwencje dla benchmarkingu. Pojedyncza liczba agregująca wyniki z wielu zadań może ukrywać odmienne profile poznawcze. Dwa modele mogą osiągać tę samą średnią, podczas gdy pierwszy jest stabilny w większości domen, a drugi oscyluje między wynikami wybitnymi a bardzo słabymi. Z punktu widzenia praktycznego te dwa systemy nie są równoważne. Niezawodność wymaga bowiem nie tylko wysokiej oczekiwanej jakości, ale również znajomości warunków, w których system zawodzi. Stanford AI Index 2026 dodatkowo zwraca uwagę, że same benchmarki coraz szybciej tracą zdolność różnicowania modeli oraz bywają obciążone błędnymi lub nieważnymi pytaniami. Raport wskazuje, że niektóre popularne zbiory testowe zawierają istotny odsetek wadliwych pozycji, a frontier models saturują nowe benchmarki znacznie szybciej, niż wcześniej przewidywano. Oznacza to, że pomiar inteligencji sztucznej coraz mniej przypomina egzamin z ustalonego zestawu pytań, a coraz bardziej wymaga ciągłego projektowania doświadczeń diagnostycznych.

Tutaj Hofstadterowska metoda „scattering experiments” odzyskuje znaczenie. Zamiast jednego kanonicznego testu należy atakować system z wielu kierunków, zmieniać formę problemu, obserwować przypadki graniczne i szukać charakterystycznego profilu błędów. Autor książki rozwija tę intuicję właśnie jako alternatywę wobec prostego sprawdzania wyników: o naturze systemu więcej może powiedzieć szeroki rozrzut zachowań niż jeden spektakularny sukces. To podejście jest bliskie współczesnej psychometrii, w której pojedyncza odpowiedź nie stanowi pełnego pomiaru cechy latentnej. Potrzebna jest rodzina zadań różniących się trudnością i strukturą.

Odporność semantyczna jako miara dojrzałej inteligencji

W przypadku AI sytuacja jest jeszcze bardziej złożona, ponieważ sama „cecha latentna” może nie być jednowymiarowa. Zamiast jednego czynnika inteligencji możemy mieć zestaw częściowo zależnych kompetencji: reprezentację relacyjną, pamięć roboczą, planowanie, kontrolę błędów, wiedzę o świecie, orientację przestrzenną, metapoznanie i zdolność korzystania z narzędzi. Model może więc przypominać nie jeden umysł o określonym IQ, lecz nietypową konfigurację wielu systemów poznawczych. Dlatego porównania typu „AI osiągnęła poziom człowieka” są naukowo nieprecyzyjne, jeśli nie wskazują populacji, rodzaju zadania, warunków testu i wymiaru kompetencji. Człowiek również nie posiada jednego „poziomu”, który można by bez reszty przenieść na wszystkie domeny.

W przypadku AI problem ten staje się jeszcze bardziej widoczny. Szczególnie istotnym przejawem poszarpania jest wrażliwość na sformułowanie zadania. Dla człowieka dwie instrukcje mogą być oczywistymi parafrazami, podczas gdy model może reagować na nie odmiennie, ponieważ zmiana powierzchni wpływa na wewnętrzne aktywacje i prawdopodobieństwa dalszych trajektorii generowania. Nie musi to oznaczać całkowitego braku rozumienia – człowiek także bywa podatny na framing.

Jednak duża skala takich efektów pokazuje, że stabilność znaczenia pozostaje odrębnym wymiarem kompetencji. Można powiedzieć, że dojrzałe pojęcie powinno wykazywać pewną odporność semantyczną: zmiana nieistotnych własności formy nie powinna radykalnie zmieniać jego funkcji. Jest to właściwie odpowiednik inwariancji z Letter Spirit. Litera powinna pozostać rozpoznawalna mimo zmiany stylu, reguła mimo zmiany narracji, a analogia mimo wymiany obiektów. Współczesna ocena inteligencji może zatem korzystać z tego samego kryterium: badać, jakie

zmiany powierzchniowe system potrafi absorbować bez utraty struktury. Jeszcze bardziej wymagające są problemy kontrfaktyczne.

Jeśli model rozumie strukturę, powinien być w stanie tymczasowo zaakceptować założenie sprzeczne z rzeczywistością i konsekwentnie wyprowadzać jego skutki. Kontrfaktyczność wymaga oddzielenia wiedzy domyślnej od lokalnie zadanej reguły. W języku FARG można powiedzieć, że system musi poluzować głęboką, dobrze utrwaloną relację, nie niszcząc przy tym globalnej koherencji nowego mikroświata.

To bardzo trudny test, ponieważ duży model jest silnie przyciągany przez regularności treningowe. Polecenie „załóżmy, że...” musi na czas lokalnego zadania pokonać ogromną wagę wcześniejszych skojarzeń. Jeśli system potrafi to zrobić konsekwentnie, demonstrowe pewien rodzaj dynamicznej kontroli reprezentacji. Jeśli po kilku krokach „wraca” do wiedzy domyślnej, ujawnia granicę kontekstowej rekonfiguracji. Zjawisko to pokazuje, że generalizacja i sterowalność są powiązane. Model inteligentny nie tylko posiada bogatą wiedzę, ale potrafi selektywnie zawieszać jej fragmenty, przyjmować nowe reguły i utrzymywać je w lokalnym kontekście. Przypomina to człowieka rozwiązującego problem matematyczny w systemie aksjomatycznym różnym od codziennej intuicji; nie musi on wierzyć, że alternatywne aksjomaty opisują świat, lecz musi potrafić konsekwentnie działać w ich obrębie. Dojrzała abstrakcja wymaga zatem rozróżnienia pomiędzy prawdą globalną, założeniem lokalnym a regułą transformacji. System, który miesza te poziomy, może generować poprawne pojedyncze odpowiedzi, lecz nie utrzyma spójnej struktury przez dłuższy ciąg rozumowania. Właśnie tutaj testy wieloetapowe mają przewagę nad pytaniami jednopunktowymi: ujawniają, czy reprezentacja pozostaje stabilna w czasie. Odnosi się to do koncepcji Metacata.

Poszarpana inteligencja nie jest problemem wyłącznie wtedy, gdy system popełnia błędy. Staje się znacznie poważniejsza, gdy system nie rozpoznaje, gdzie jego kompetencja jest poszarpana. Gdyby model potrafił niezawodnie stwierdzić: „w tej klasie zadań moja generalizacja jest krucha”, część ryzyka zostałaby zredukowana. Najbardziej niebezpieczna konfiguracja pojawia się wtedy, gdy duża zdolność generatywna współistnieje z niewystarczającą kalibracją własnych ograniczeń. Ostatecznie więc granica między imitacją analogii a funkcjonalną analogią nie przebiega w jednym punkcie – jest wielowymiarowa.

Należy pytać o transfer, odporność na perturbacje, kontrfaktyczność, kompozycjonalność, generalizację poza rozkład, stabilność w czasie, profil błędów i metapoznawczą ocenę własnego wyniku. Dopiero łączna analiza tych wymiarów pozwala sensownie mówić o stopniu, w jakim system posiada funkcjonalnie płynne reprezentacje. To rozstrzygnięcie prowadzi do ważnego wniosku filozoficznego: inteligencja nie jest po prostu zdolnością do udzielenia odpowiedzi, lecz

zdolnością do zachowania właściwej struktury podczas zmian, które nie powinny mieć znaczenia, i do zmiany struktury wtedy, gdy zmienia się to, co znaczenie mieć powinno. W pierwszym przypadku potrzebna jest stabilność, w drugim – plastyczność. System pozbawiony stabilności jest chaotyczny; system pozbawiony plastyczności jest sztywny. Płynność Hofstadtera okazuje się więc nie synonimem zmienności, ale zdolnością regulowania relacji pomiędzy trwałością a zmianą.

Od imitacji do diagnostyki struktur poznawczych

Współczesne modele językowe znajdują się właśnie w tej przestrzeni pośredniej. Ich sukcesy dowodzą, że potrafią budować i wykorzystywać abstrakcje na poziomie, którego nie da się wyjaśnić prostą metaforą wyszukiwania gotowych zdań. Porażki pokazują jednak, że abstrakcje te nie zawsze posiadają stabilność właściwą ludzkim systemom pojęciowym. Kompetencja bywa niezwykle głęboka lokalnie, lecz jej granice mogą przebiegać w miejscach nieintuicyjnych. Nie oznacza to, że człowiek dysponuje doskonale spójnym systemem pojęć – psychologia od dawna wskazuje na kontekstowość kategorii, błędy framingu, podatność na priming, interferencję pamięci czy heurystyczność rozumowania. Różnica tkwi raczej w historii stabilizowania pojęć poprzez ucieleśnienie, działanie i praktykę wspólnotową. Człowiek przez lata wiąże słowa z percepcją, ruchem, konsekwencjami i społecznym sprzężeniem zwrotnym. Model językowy przejmuje część tej struktury pośrednio z danych, a nowsze systemy multimodalne i agentowe stopniowo poszerzają zakres bezpośredniego sprzężenia.

Dlatego pytanie o przyszłość nie powinno brzmieć, czy kolejna generacja modeli „wreszcie osiągnie prawdziwą analogię”. Bardziej naukowe jest pytanie: które typy inwariancji, transferu i samokontroli będą stopniowo stabilizowane wraz z rozwojem architektur, treningu, pamięci, multimodalności i działania?

Możemy spodziewać się przesuwania jagged frontier, ale nie ma podstaw zakładać, że będzie ona znikać równomiernie lub według jednego, przewidywalnego porządku. Przyszłość wymyka się zatem prostej predykcji. Historia sztucznej inteligencji wielokrotnie pokazywała, że przełomy zmieniają nie tylko poziom wyników, ale i strukturę pytań. Systemy, które miały być modelami inteligencji, okazywały się zbyt wąskie; z kolei predyktory językowe uzyskiwały kompetencje, których skala zaskoczyła znaczną część środowiska. Ekstrapolowanie jednej krzywej wydajności nie wystarcza, gdyż nowe architektury i metody treningu mogą całkowicie reorganizować profil zdolności. Z tego punktu widzenia „jagged intelligence” nie jest chwilową usterką obecnej generacji AI, lecz przede wszystkim ostrzeżeniem epistemologicznym przed zbyt szybkim budowaniem liniowej hierarchii inteligencji. Być może nie istnieje jedna drabina, po której maszyna wspina się

od „głupoty” do „ludzkiego poziomu”, lecz wielowymiarowy krajobraz zdolności, w którym system może zajmować pozycje niemające odpowiednika w populacji ludzkiej. Takie spojrzenie wymusza ponowny namysł nad Testem Turinga.

Jeśli system może być bardziej kompetentny od człowieka w jednych domenach i słabszy w innych, a zarazem perfekcyjnie naśladować ludzką rozmowę, sama nierozróżnialność konwersacyjna staje się bardzo słabą mapą jego architektury poznawczej. Test Turinga pozostaje historycznie genialnym eksperymentem epistemologicznym, ale współczesne systemy wymagają znacznie szerszej diagnostyki. Zamiast jednego pytania „czy rozmówca jest człowiekiem?”, potrzebujemy całego układu prób: czy system utrzymuje inwariant przy zmianie powierzchni, czy przenosi relację między domenami, czy rozpoznaje własny błąd, czy zachowuje kontrfaktyczną regułę, czy potrafi zrekonstruować własne ograniczenia, czy jego kompetencja wytrzymuje presję nowych modalności oraz czy wynik jest skutkiem stabilnej struktury, czy kruchego dopasowania.

W tym miejscu Hofstadterowskie „scattering experiments” przestają być historyczną ciekawostką, a stają się zapowiedzią współczesnego programu ewaluacji. Zamiast jednego ceremonialnego egzaminu inteligencji należałoby wystawiać system na problemy o różnej strukturze, odległości od danych treningowych i wymaganym typie reprezentacji. Dopiero rozkład sukcesów, porażek, reakcji na impas i zachowania po zmianie kontekstu dawałby poznawczo wartościowy portret. To właśnie będzie przedmiotem kolejnego etapu wywodu.

Skoro pojedyncza rozmowa coraz słabiej nadaje się do rozstrzygania natury systemu, trzeba powrócić do Testu Turinga z perspektywy doświadczeń związanych z LLM. Nie po to, by symbolicznie ogłosić, że maszyna „zdała” egzamin na człowieczeństwo, lecz aby zapytać: co właściwie powinniśmy mierzyć, gdy imitacja językowa przestała być wystarczająco trudnym kryterium? Wtedy Test Turinga przekształci się z pojedynczej gry imitacyjnej w szerszy problem epistemologiczny: jak odróżnić zachowanie przekonujące od mechanizmu poznawczo głębokiego i jakie eksperymenty pozwalają badać tę różnicę bez popadania ani w antropomorfizm, ani w redukcjonizm.

Być może nie istnieje jedna drabina, po której maszyna wspina się od prymitywnych obliczeń do ludzkiego poziomu inteligencji, lecz wielowymiarowy krajobraz zdolności, w którym systemy AI zajmują pozycje całkowicie obce naszej naturze. W tym świecie Test Turinga przestaje być ostatecznym egzaminem z człowieczeństwa, a staje się jedynie powierzchowną grą pozorów. Prawdziwym wyzwaniem nie jest już pytanie, czy maszyna

potrafi nas przekonać, że myśli, lecz czy potrafimy zaprojektować eksperymenty zdolne odróżnić głęboką strukturę poznawczą od perfekcyjnej imitacji.

Inspiracje

[1]



"Fluid Concepts and Creative Analogies" Douglas Hofstadter

ISBN: 9780140258356

Douglas Richard Hofstadter (ur. 15 lutego 1945 r.) jest amerykańskim uczonym i wybitnym profesorem kognitywistyki i literatury porównawczej na Uniwersytecie Indiany w Bloomington. Specjalizuje się w kognitywistyce, sztucznej inteligencji, filozofii umysłu oraz naturze ludzkiego myślenia i kreatywności. Hofstadter jest uznawany na całym świecie za pionierską pracę badawczą nad autoreferencyjnością, analogią, poznaniem i świadomością. Zdobył szerokie uznanie książką z 1979 roku „Gödel, Escher, Bach: An Eternal Golden Braid”, która otrzymała Nagrodę Pulitzera w kategorii literatury faktu oraz Amerykańską Nagrodę Książkową w dziedzinie Nauki. Badania Hofstadtera wywarły głęboki wpływ na modelowanie poznawcze, szczególnie poprzez architektury obliczeniowe badające tworzenie analogii, formowanie pojęć i wyłaniającą się naturę „ja” jako dziwnej pętli.

Douglas Richard Hofstadter (born February 15, 1945) is an American scholar and Distinguished Professor of Cognitive Science and Comparative Literature at Indiana University Bloomington. He specializes in cognitive science, artificial intelligence, philosophy of mind, and the nature of human thought and creativity. Hofstadter is internationally recognized for his pioneering work exploring self-reference, analogy, cognition, and consciousness. He achieved widespread acclaim with his 1979 book Gödel, Escher, Bach: An Eternal Golden Braid, which was awarded the Pulitzer Prize for General Nonfiction and the American Book Award for Science. Hofstadter's research has profoundly influenced cognitive modeling, particularly through computational architectures exploring analogy-making, concept formation, and the emergent nature of the self as a strange loop.

Indiana University Bloomington

Literatura uzupełniająca

Książki

Książki autora i dzieła przywołane w tekście:



[1] "The Mind's I"

Douglas Hofstadter

Hachette UK | ISBN: 9780786725144

[2] "Trying to Muse Rationally About the Singularity Scenario"

Douglas Hofstadter



[3] "Gödel, Escher, Bach"

Douglas Hofstadter

1979 | ISBN: 9780140179972

[4] "The Mind's I: Fantasies And Reflections On Self & Soul"

Douglas Hofstadter



[5] "Metamagical Themas"

Douglas R Hofstadter

2008 | Hachette UK | ISBN: 9780786723867

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] "Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data"

Andrea Gregor de Varda, Marco Marelli, Simona Amenta

2023 | Behavior Research Methods | DOI: 10.3758/s13428-023-02261-8

[Źródło](#)

[2] "Scaling in Cognitive Modelling: a Multilingual Approach to Human Reading Times"

Andrea Gregor de Varda, Marco Marelli

2023 | Proceedings of the conference - Association for Computational Linguistics. Meeting | DOI: 10.18653/v1/2023.acl-short.14

[Źródło](#)

[3] "The Effects of Surprisal across Languages: Results from Native and Non-native Reading"

Andrea Gregor de Varda, Marco Marelli

2022 | Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022 | DOI: 10.18653/v1/2022.findings-acl.13

[Źródło](#)

[4] "Meaning Beyond Lexicality: Capturing Pseudoword Definitions with Language Models"

Andrea Gregor de Varda, Daniele Gatti, Marco Marelli, Fritz Günther

2024 | Computational Linguistics | DOI: 10.1162/coli_a_00527

[Źródło](#)

Artykuły naukowe autora:

[5] "Godel, Escher, Bach: An Eternal Golden Braid."

Jonathan Lieberman, Douglas R. Hofstadter

1980 | The Journal of Philosophy | DOI: 10.2307/2025598

[Źródło](#)

[6] "Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought"

Douglas R. Hofstadter

1995

[Źródło](#)

[7] "Metamagical Themas: Questing for the Essence of Mind and Pattern"

Charles Taylor, Douglas R. Hofstadter

1985 | The Antioch Review | DOI: 10.2307/4611504

[Źródło](#)

[8] "Fluid Concepts and Creative Analogies"

Douglas R. Hofstadter

1995

[Źródło](#)

[9] "Epilogue: Analogy as the Core of Cognition"

Douglas R. Hofstadter

2001 | The MIT Press eBooks | DOI: 10.7551/mitpress/1251.003.0020

[Źródło](#)

[10] "High-level perception, representation, and analogy: A critique of artificial intelligence methodology"

David J. Chalmers, Robert M. French, Douglas R. Hofstadter

1992 | Journal of Experimental & Theoretical Artificial Intelligence | DOI: 10.1080/09528139208953747

[Źródło](#)

[11] "The copycat project: A model of mental fluidity and analogy-making"

Douglas R. Hofstadter, Melanie Mitchell

1995

[Źródło](#)

[12] "The Mind's I: Fantasies and Reflections on Self and Soul."

Andrew Hamilton, Douglas R. Hofstadter, Daniel C. Dennett

1984 | The Philosophical Quarterly | DOI: 10.2307/2218896

[Źródło](#)



Treść tworzy Fundacja Dobre Państwo.

Redaguje i publikuje APA ONE, autorski system redakcyjny Fundacji oparty na AI.

APA ONE Publishing System

Fundacja Dobre Państwo © 2026

Article ID: dc050077-2cd4-4e9e-8167-c2f3a96c3de8