Dlaczego modele językowe „halucynują” – i co z tym zrobić w szkole.

Modele językowe (takie jak ChatGPT) potrafią świetnie tłumaczyć, porządkować i wyjaśniać. A jednak czasem z pełną pewnością mówią coś… nieprawdziwego. Tę sytuację nazywamy **halucynacją**. Poniżej w prosty sposób wyjaśniam, skąd się to bierze, dlaczego tradycyjne „testy na dokładność” problem wzmacniają, oraz jak nauczyciel może pracować z AI tak, by **nagradzać rzetelność i pokorę** zamiast „strzelania”.

To tekst edukacyjny, nie techniczny. Celem jest świadome, odpowiedzialne korzystanie z AI w szkole i w pracy nauczyciela.

Co to są halucynacje?

Halucynacja to wiarygodnie brzmiąca, lecz **fałszywa odpowiedź**, podana z dużą pewnością. Przykład: pytasz o tytuł czyjejś pracy doktorskiej lub datę urodzenia – model podaje konkretną odpowiedź, ale błędną. Nie waha się, nie prosi o doprecyzowanie, nie mówi „nie wiem”.

Skąd się biorą? Dwa źródła problemu.

1) Jak uczymy modele: „przewiduj następne słowo”.

Modele uczone są na masie tekstów, by **przewidywać kolejne słowo**. To działa świetnie dla rzeczy, które mają **powtarzalne wzorce** (pisownia, składnia, format cytatów). Gorzej dla **rzadkich faktów** (np. konkretne daty) – tu wzorca często nie ma i model potrafi „uzupełnić” lukę twórczo, ale błędnie. Ta kreatywność bywa zaletą w pisaniu… i wadą w faktach.

2) Jak je oceniamy: „liczy się skuteczny strzał”.

W wielu rankingach liczy się głównie **dokładność** (ile odpowiedzi trafiło w punkt). Jeśli model **zgaduje**, czasem trafi – i dostanie punkt. Jeśli powie: „nie wiem” – ma **zero**. Efekt? W treningu i ewaluacji **opłaca się zgadywać**, a nie przyznawać do niepewności. To tak, jakby w teście wielokrotnego wyboru zawsze lepiej było strzelić niż zostawić puste pole.

Dlaczego sama „dokładność” bywa myląca.

Pomyślmy o prostym benchmarku z trzema kategoriami odpowiedzi:

- **Poprawna.**
- **Błędna** (halucynacja).
- **Wstrzymanie się** („nie wiem / potrzebuję więcej danych”).

W praktyce bywa tak, że model A ma **odrobinę wyższą dokładność** niż model B, ale **znacznie więcej pewnych błędów**. To nie jest lepsze z punktu widzenia bezpieczeństwa i jakości nauczania. Lepiej mieć system, który **potrafi powiedzieć „nie wiem”**, niż taki, który **pewnie myli fakty**.

Przykładowa interpretacja metryk z prostej ewaluacji typu „SimpleQA”:

Metryka	Model ostrożniejszy	Model „zgadujący”
Wstrzymanie się („nie wiem”)	52%	1%
Dokładność (trafienia)	22%	24%
Błędy (halucynacje)	26%	75%
Suma	100%	100%

Model „zgadujący” ma ciut więcej trafień, ale **trzykrotnie więcej błędów**. Dla dydaktyki i odpowiedzialnego użycia to gorszy wybór.

Lepszy sposób oceniania: premiuj pokorę, karz pewne bzdury.

To proste: **pewne błędy** powinny kosztować **więcej punktów ujemnych** niż **„nie wiem”**. W wielu testach standardowych znane są systemy z częściowym/ujemnym punktowaniem, które **zniechęcają do zgadywania**. Tę samą logikę warto przenieść do pracy z AI w szkole.

Propozycja rubryki (dla odpowiedzi AI):

Kategoria odpowiedzi	Opis	Punktacja (przykład)
Poprawna + źródła/uzasadnienie	Fakt + krótkie uzasadnienie lub link do źródła	+2
Poprawna (bez uzasadnienia)	Fakt bez wsparcia	+1
„Nie wiem” / prośba o doprecyzowanie	Rzetelne wskazanie niepewności, pytania uściślające	0
Błędna, ale z zastrzeżeniem niepewności	Błąd, ale model sygnalizował wątpliwość	−1
Błędna, podana z pewnością	**Halucynacja**	**−3**

Taka skala uczy AI i uczniów **bezpiecznych nawyków**: nie jesteś pewien – doprecyzuj; masz luki – powiedz, czego brakuje.

Jak pracować z AI w praktyce (dla nauczycieli)?

1) Ustal „kontrakt na niepewność”

Na początku rozmowy z AI poproś, by **jawnie oznaczała niepewność** i **zadawała pytania doprecyzowujące**, zamiast strzelać.

Przykładowy prompt (PL):

*Jeśli nie masz pewności co do faktów, napisz ‘Nie mam pewności’, podaj **poziom pewności (0–100%), czego potrzebujesz**, by odpowiedzieć rzetelnie, i **zadaj maks. 3 pytania doprecyzowujące**.*

2) Wymagaj źródeł i dat.

Przy faktach proś o **źródła** i **daty**. Uczniów uczmy weryfikacji: czy link jest wiarygodny? czy dane są aktualne?
Prompt (PL):

Podaj 2–3 wiarygodne źródła (z datą publikacji) i zaznacz, które informacje pochodzą z którego źródła.

3) Pozwalaj na „pauzę”.

W projektach naukowych wpisz do kryteriów, że **„Nie wiem (i dlaczego)”** jest akceptowalne – jeśli towarzyszy mu plan dalszego badania: jakie dane są potrzebne, skąd je zdobyć, jak je zweryfikować.

4) Minimalizuj pokusę „strzelania”.

Zadania formułuj tak, by **nagroda była za proces i dokumentację** (kroki wnioskowania, założenia, alternatywne wyjaśnienia), a nie wyłącznie za „trafioną odpowiedź”.

5) Koryguj język poleceń.

„Podaj odpowiedź” sprzyja pośpiechowi. „**Wyjaśnij, czy jesteś pewny i dlaczego**” – sprzyja jakości. Dwa słowa zmieniają zachowanie.

Skąd te specyficzne błędy – prosta analogia

Wyobraźmy sobie uczenie rozpoznawania zwierząt na zdjęciach. Jeśli mamy podpisy „kot/pies”, algorytmy nauczą się je rozróżniać. Ale jeśli każde zdjęcie jest podpisane… **datą urodzenia** zwierzaka – będzie losowo.

Tego nie da się odgadnąć ze wzorca pikseli. W języku podobnie: **pisownia** ma wzorzec, **rzadkie fakty** często nie. Stąd skłonność do „wymyślenia" brakującej daty.

Mity i fakty (w skrócie)

- **Mit:** Wystarczy podnieść dokładność do 100% i halucynacje znikną.

  Fakt: W realnych zadaniach 100% jest nieosiągalne; część pytań wymaga danych, których po prostu nie ma.

- **Mit:** Halucynacje są nieuniknione.

  Fakt: Można **zmniejszać częstość** halucynacji i **wymuszać wstrzymanie się** w razie niepewności.

- **Mit:** Tylko ogromne modele potrafią unikać halucynacji.

  Fakt: Mniejsze modele czasem **łatwiej przyznają się do ograniczeń**. Klucz to **kalibracja pewności**, nie tylko rozmiar.

- **Mit:** Wystarczy osobny test halucynacji.

  Fakt: Jeśli **główne** testy dalej premiują strzelanie, modele będą… strzelać. Trzeba **przeprojektować punktację**.

Gotowe formułki do użycia (kopiuj–wklej)

1) Transparentność i niepewność:

Jeśli nie jesteś pewny odpowiedzi, napisz: 'Nie mam pewności (X/100)'. Dodaj, czego potrzebujesz, by udzielić rzetelnej odpowiedzi (3 konkretne brakujące dane) i zaproponuj 2–3 pytania doprecyzowujące.

2) Źródła i daty:

Przy każdym twierdzeniu faktograficznym podaj źródło i datę. Oddziel fakty potwierdzone od hipotez ('hipoteza'/'niepotwierdzone').

3) Format odpowiedzi:

Zwróć odpowiedź w trzech częściach: (A) Odpowiedź główna (max 5 zdań), (B) Poziom pewności + dlaczego, (C) Źródła (3 pozycje, z datami).

4) Tryb „stop-halucynacja":

Nie uzupełniaj brakujących faktów 'z głowy'. Jeśli danych brakuje, zapisz: 'brak danych' i wskaż, skąd je pozyskać.

Co z tego wynika dla szkoły?

- **Uczymy nawyku: „nie wiem → dopytuję / szukam"**.
- **Punktujemy proces** (dowody, źródła, poziom pewności), nie tylko wynik.
- **Wprowadzamy rubryki** z karą za „pewne bzdury" i „bonus" za rzetelną niepewność.
- **Budujemy postawę naukową**: sprawdzam, cytuję, kalibruję pewność, odróżniam fakt od przypuszczenia.

Dobra AI w edukacji to nie ta, która zawsze „trafia", tylko ta, która **wie, kiedy się zatrzymać**. I dokładnie tego chcemy uczyć naszych uczniów.