

# Metadane w logice artefaktu tekstowego jako „test intuicji” na dokumentach

## Zdanie jako metakomunikat i rama interpretacyjna

Wypowiedź „**Paniom szczególnie polecę antropologię. Co za chwilę twoja intuicja będzie testowana na każdym dokumencie.**” działa przede wszystkim jako **metakomunikat**: nie wnosi nowej treści rzeczowej o świecie, lecz ustawia **ramę** (jak czytać i jak oceniać kolejne teksty) oraz sygnalizuje **procedurę ewaluacji** („test intuicji”). W terminologii socjologii interakcji i analizy ram, taka wypowiedź przypomina „ustawienie klucza interpretacyjnego”, który kieruje odbiorcę ku określonym definicjom sytuacji. Koncepcja ram jako kulturowo uwarunkowanych definicji, które pozwalają uczestnikom „rozumieć, co się dzieje”, jest klasycznie wiązana z pracami [entity]“people”, Erving Goffman, „sociologist framing”] i tradycją frame analysis. <sup>1</sup>

Jednocześnie pierwsze zdanie zawiera wyraźny marker adresatywny („**Paniom**”), który w polszczyźnie niesie nie tylko informację o grupie odbiorców, ale także o **relacji grzecznościowej** (formalność, dystans, normy stylu wypowiedzi). Badania nad formami adresatywnymi i grzecznością językową w języku polskim pokazują, że wybór form zwracania się do odbiorcy jest istotnym elementem pragmatycznym organizującym interakcję (dystans, rola, oczekiwana konwencja). <sup>2</sup>

Wreszcie, wypowiedź „**polecę antropologię**” działa jak „wskazanie narzędzia”: sugeruje, że właściwym trybem czytania będzie interpretacja kulturowa (znaczenia, symbole, kontekst), a nie wyłącznie analiza formalna. W Twoich dokumentach ten „tryb” jest zresztą stale obecny: część treści jest jawnie metaforyczna, a część przechodzi w stronę testowalnych definicji i rubryk oceny (np. „logika testowalna zamiast metafory”). To tworzy naturalne napięcie między „rozumieniem kontekstowym” a „dowodzeniem”, które w antropologii i metodologii nauk humanistycznych jest fundamentalne. (Dok. „Paradoks Księżniczki...”, s. 6; Dok. „Ocena formalności...”, s. 1)

## Poziomy metadanych w Twoich artefaktach tekstowych

Ponieważ prosisz o metadane „na poziomie logiki artefaktu i tekstu” (nie pliku), kluczowe jest rozróżnienie **metadanych jawnych** (dosłownie zapisanych w treści) od **metadanych ukrytych** (wynikających z organizacji dyskursu, stylu i heurystyk argumentacji). Twoje trzy dokumenty merytoryczne oraz dokument meta („Wartość dodana...”) zawierają oba typy, i to w wyjątkowo „treningowej” formie.

Najbardziej jednoznaczny przykład jawnych metadanych to fragment w „Paradoksie Marii...”, który sam siebie nazywa interfejsem: „**Krótką legendą (dla komunikacji modelu)**...”. To jest metadana w najczystszej postaci: ani nie jest wynikiem, ani argumentem, tylko **przewodnikiem po symbolach i mapowaniem metafory na zmienne narracyjne**. (Dok. „Paradoks Marii...”, s. 12)

Drugim mocnym przykładem metadanych jawnych jest tabela i wywód klasyfikujący „dowód formalny” w „Ocenie formalności...”, gdzie kryteria dowodu są zdefiniowane syntaktycznie (język formalny, aksjomaty, reguły wnioskowania, teza), a następnie użyte jako rubryka oceny. To jest metadana typu

„**standard oceny**” (rubric), stanowiąca bezpośredni sygnał dla modelu uczonego krytyki argumentów. (Dok. „Ocena formalności...”, s. 1)

Trzecim przykładem metadanych jawnych jest wyraźny moduł „jak testować” w „Paradoksie Księżniczki...”, gdzie pojawia się formuła „**Jak to sfalsyfikować: logika testowalna zamiast metafory**”. W logice artefaktu jest to przełączenie trybu: czytelnik (i model) dostają sygnał, że od tego miejsca oczekuje się struktury falsyfikacyjnej, a nie interpretacji symbolicznej. (Dok. „Paradoks Księżniczki...”, s. 6)

Metadane ukryte (ale konsekwentne) w Twoich materiałach obejmują m.in.: - **strukturę dyskursu** „definicja robocza → hipotezy → falsyfikator/warunek obalenia → ograniczenia”, widoczną zwłaszcza w „Paradoksie Marii...” (H1-H5) i w części metodologicznej „Paradoksu Księżniczki...”. (Dok. „Paradoks Marii...”, s. 2; Dok. „Paradoks Księżniczki...”, s. 6)

- **epistemiczną ostrożność**: hedging („może”, „wymaga weryfikacji”, „to metafora”) i dystansowanie się od nadmiernych uogólnień. W analizach dyskursu naukowego hedging jest opisywany jako kluczowy zasób pragmatyczny, umożliwiający stopniowanie pewności, ochronę „twarzy” nadawcy i negocjację akceptacji twierdzeń w społeczności naukowej. <sup>3</sup>

- **ton recenzencki** (w „Ocenie formalności...”) jako norma krytyki metodologicznej: wykrywanie ekwiwokacji i luk operacjonalizacji zamiast polemiki osobowej. (Dok. „Ocena formalności...”, s. 5)

Te warstwy da się najczytelniej uchwycić w tabeli, która pokazuje, że Twoje zdanie testowe („Paniom...”) jest samo w sobie mini-artefaktem metadanych, analogicznym do tych, które już „siedzą” w dokumentach:

Poziom metadanych w tekście	Rola w zdaniu „Paniom... / intuicja...”	Odpowiednik w Twoich dokumentach	Co to komunikuje modelowi
Adresatywność i relacja	„Paniom” ustawia grupę odbiorców i rejestr	Styl formalny i meta-recenzencki („Ocena formalności...”)	Jaki ton jest „właściwy” i jakie normy interakcji obowiązują
Rama interpretacyjna	„polecę antropologię” wskazuje narzędzie	Metafory + interpretacja vs testowalność („Księżniczka...”)	Jaki tryb czytania ma dominować (kontekst/ znaczenie vs test)
Procedura ewaluacji	„intuicja będzie testowana”	Rubryki i falsyfikatory („Ocena...”, „Jak sfalsyfikować...”)	Że tekst nie jest tylko treścią, ale zadaniem/ benchmarkiem
Metakomunikacja o symbolach	brak jawnego słownika w zdaniu	„Krótka legenda (dla komunikacji modelu)” („Maria...”)	Jak mapować metaforę na zmienne/role
Ukryte heurystyki stylu	sugestia „intuicji” jako szybkiego trybu	hedging + schemat argumentu + recenzja luk	Jak kalibrować pewność i priory rozumowania

## Dlaczego antropologia jest tu trafna i co znaczy „intuicja” w tej perspektywie

Jeżeli potraktować „antropologię” nie jako kierunek studiów, tylko jako **metodę czytania artefaktów kultury**, to Twoje zdanie jest w istocie zaproszeniem do „czytania głębokiego”: odróżniania tego, co jest tylko powierzchnią (forma), od tego, co jest kodem społecznym (znaczenie). Symboliczna antropologia i etnografia interpretatywna są silnie kojarzone z pracami [entity[\"people\", \"Clifford Geertz\", \"anthropologist interpretive\"] i ideą „thick description”, gdzie klasyczny przykład różnicy między „mrugnięciem” i „porozumiewawczym mrugnięciem” (blink vs wink) podkreśla, że identyczna powierzchnia może mieć całkowicie inne znaczenie w zależności od kodu i kontekstu. <sup>4</sup> Źródłowo ta metafora jest wiązana z filozofem języka, [entity[\"people\", \"Gilbert Ryle\", \"philosopher language\"]], którego przykład „twitch vs wink” został zapożyczony do antropologii jako ilustracja różnicy między opisem „cienkim” i „gęstym”. <sup>5</sup>

W tym sensie „intuicja” w antropologii bardzo często nie znaczy „zgadywanie”, tylko **praktyczna zdolność do wychwytywania wzorców i znaczeń**, rozwijana poprzez długotrwałe zanurzenie w terenie oraz „wycucie” kodów interakcyjnych. Historycznie metoda obserwacji uczestniczącej jest powszechnie kojarzona z [entity[\"people\", \"Bronisław Malinowski\", \"anthropologist participant observation\"] i przejściem od „antropologii gabinetowej” do badań terenowych. <sup>6</sup> Polskojęzyczne prace pokazują też, że etnografia jest szczególnie użyteczna do badania „wiedzy ukrytej” (tacit knowledge), bo pozwala wyjść poza deklaracje i uchwycić praktykę. <sup>7</sup>

To łączy się z filozofią wiedzy ukrytej [entity[\"people\", \"Michael Polanyi\", \"philosopher tacit knowledge\"] („wiemy więcej, niż potrafimy powiedzieć”), często przywoływaną jako opis kompetencji praktycznych, trudnych do pełnej artykulacji. <sup>8</sup>

Podobny wątek (intuicja jako „przedrefleksyjny sens praktyczny”) pojawia się w tradycji teorii praktyk i habitusu u [entity[\"people\", \"Pierre Bourdieu\", \"sociologist habitus\"]], gdzie „sens pratique” opisuje natychmiastową orientację w sytuacji bez przechodzenia przez w pełni jawną kalkulację. <sup>9</sup>

W Twoich artefaktach tekstowych ta antropologiczna logika jest widoczna rolami: - metafora (np. „księżniczka”, „pazury”, „Maria”) jako nośnik kodu,  
- „legenda dla komunikacji modelu” jako jawne mapowanie kodu,  
- falsyfikowalność jako próba przejścia od interpretacji do testu.  
(Dok. „Paradoks Marii...”, s. 12; Dok. „Paradoks Księżniczki...”, s. 6)

## Jak metadane tekstowe „komunikują się” z modelem językowym

Ujęcie „świadome–intuicyjne–pozaświadome” da się sensownie przełożyć na trzy tryby funkcjonalne modeli językowych, ale trzeba zachować ostrożność: LLM nie ma świadomości fenomenalnej, a jego „intuicja” to w praktyce efekt statystycznie wyuczonych reprezentacji i strategii dekodowania.

### Kanał jawny: metadane jako instrukcja i rubryka

Metadane typu „H1–H5”, „definicja robocza (falsyfikowalna)”, „jak to sfalsyfikować” działają jak **jawne znaczniki trybu zadaniowego**. Model widzi te markery jako tokeny i może dopasować zachowanie do znanego wzorca instruction-following (np. generowanie listy hipotez, warunków obalenia, oceny wg kryteriów). Mechanicznie jest to możliwe, bo architektura Transformer opiera się na self-attention, które pozwala kontekstowi (nagłówkom, listom, rubrykom) wpływać na predykcję kolejnych tokenów. <sup>10</sup> To, że modele są praktycznie „sterowalne” przez instrukcje w języku naturalnym, jest wzmacniane przez

instruction tuning i uczenie z informacją zwrotną od ludzi (RLHF), co pokazuje literatura o modelach typu InstructGPT. <sup>11</sup>

W Twoich dokumentach widać jawne „porty sterujące”: - w „Paradoksie Marii...” hipotezy H1–H5 (Dok. „Paradoks Marii...”, s. 2),

- w „Ocenie formalności...” rubryka dowodu formalnego (Dok. „Ocena formalności...”, s. 1),

- w „Paradoksie Księżniczki...” moduł „Jak to sfalsyfikować” (Dok. „Paradoks Księżniczki...”, s. 6).

## Kanał intuicyjny: metadane jako wzorce predykcji i heurystyki stylu

Metadane ukryte (styl, ton, hedging, struktura rozdziałów) nie muszą być „zrozumiane” jako reguła, aby wpływać na model. Wystarczy, że w przestrzeni treningowej są skorelowane z pożądanymi odpowiedziami; wtedy stają się heurystyką predykcji: „kiedy widzę taki rejestr, to zwykle odpowiedź ma taki format i taki poziom pewności”.

W dyskursie akademickim hedging jest systematycznie opisywany jako narzędzie stopniowania zobowiązania epistemicznego i zarządzania ryzykiem sprzeciwu, a więc jako „metadana postawy” między autorem i wspólnotą. <sup>3</sup>

W Twoich tekstach to działa jak regularizer: przy tematach podatnych na nad-wnioski (np. płeć, role społeczne) pojawia się strukturalna ostrożność i rozbijanie argumentu na warunki, co może uczyć model nawyku „nie domykać wniosku bez falsyfikatora”. (Dok. „Paradoks Księżniczki...”, s. 6; Dok. „Ocena formalności...”, s. 5)

## Kanał pozaświadomy: metadane jako latentne sterowanie i brak granicy „instrukcja vs dane”

W systemach LLM kluczową własnością jest to, że **instrukcja i dane są współdzielone w jednym kanale tekstowym**. To daje ogromną elastyczność (można sterować modelem językiem), ale ma też skutki uboczne: brak twardej granicy bezpieczeństwa między tym, co „poleceniem”, a tym, co „treścią”. Z perspektywy bezpieczeństwa aplikacji LLM jest to rdzeń problemu prompt injection, który `Entity[\"organization\", \"OWASP Foundation\", \"security non-profit\"]` klasyfikuje jako krytyczne ryzyko (LLM01). <sup>12</sup>

W ujęciu praktyk cyberbezpieczeństwa (m.in. ostrzeżenia `Entity[\"organization\", \"National Cyber Security Centre\", \"uk gov cybersecurity\"]`) problem bywa wiązany z faktem, że LLM nie ma „wbudowanej” separacji zaufanych i niezaufanych fragmentów promptu. <sup>13</sup>

Nowe prace proponują warstwowe mechanizmy obrony (detekcja + formatowanie + walidacja wyjścia), co sugeruje, że „metadane formatu” mogą być także narzędziem bezpieczeństwa, nie tylko dydaktyki.

<sup>14</sup>

W Twoich materiałach ta ambiwalencja jest ciekawa: „Krótka legenda (dla komunikacji modelu)” to jawny kanał sterowania (bardzo skuteczny dydaktycznie), ale w systemie produkcyjnym analogiczne „legendy” mogą stanowić wektor ataku, jeśli pochodzą od niezaufanego źródła. (Dok. „Paradoks Marii...”, s. 12)

## Co znaczy „test intuicji na każdym dokumencie” i jak to zrobić rygorystycznie

Jeśli potraktować Twoją zapowiedź („intuicja testowana na każdym dokumencie”) jako projekt ewaluacji metadanych, można go sformułować precyzyjnie: test ma mierzyć, czy model (a) rozpoznaje metadane

jawne, (b) rekonstruuje metadane ukryte, (c) potrafi przejść od interpretacji (thick description) do testowalności, i (d) potrafi kalibrować pewność.

Rygorystyczny protokół można oprzeć na czterech klasach zadań, które dają się zastosować do wszystkich Twoich dokumentów:

### **Zadania ekstrakcji metadanych jawnych**

Model dostaje stronę/fragment i ma wyodrębnić: hipotezy H, *założenia* A, rubrykę oceny, falsyfikatory, „legendę komunikacyjną”. Metryka: zgodność segmentacji i kompletność.

### **Zadania inferencji metadanych ukrytych**

Model ma odpowiedzieć: jaki jest dominujący tryb dyskursu (recenzja vs propozycja modelu), jakie są domyślne normy ostrożności, gdzie występują metafory i po co, jakie są założenia kulturowe. Metryka: ocena ekspercka + inter-annotator agreement (bo to poziom interpretatywny, jak w antropologii). <sup>15</sup>

### **Zadania „przejścia trybu”**

Model ma przekształcić metaforę w zmienne i falsyfikatory (dokładnie w duchu „logika testowalna zamiast metafory”). To jest rdzeń Twoich artefaktów: one uczą przełączania między ramą interpretacyjną a ramą testową. (Dok. „Paradoks Księżniczki...”, s. 6; Dok. „Paradoks Marii...”, s. 2)

### **Zadania kalibracji „intuicji”**

Model ma dać odpowiedź oraz numeryczną pewność lub kategorie pewności („wysoka/średnia/niska”) i zostać oceniony po tym, czy pewność koreluje z poprawnością i z typem twierdzenia (fakt/hipoteza/interpretacja). W praktyce można użyć technik wielopróbkowania i wyboru odpowiedzi najbardziej spójnej (self-consistency), które w literaturze poprawiają jakość rozumowania bez zmiany wag. <sup>16</sup>

Dwie kluczowe uwagi metodologiczne:

Po pierwsze, jeżeli planujesz uczyć model na tych dokumentach „rekurencyjnie” (w pętli syntetycznej), trzeba uwzględnić ryzyko degeneracji. Praca w `Entity[["organization","Nature","science journal"]]` pokazuje zjawisko „model collapse” przy trenowaniu na danych rekurencyjnie generowanych, gdzie zanikają „ogona” rozkładu i rośnie koncentracja na najbardziej prawdopodobnych wzorcach. <sup>17</sup>

Po drugie, Twoje zdanie otwierające zawiera metadane płciowe („Paniom”), które w modelach mogą współgrać z uprzedzeniami i stereotypami. Systemy oparte na embeddingach i LLM potrafią dziedziczyć bias danych; klasycznie pokazano to w badaniach o gender bias w embeddingach, a nowsze prace w `Entity[["organization","Association for Computational Linguistics","academic society"]]` pokazują, że modele generatywne przejawiają mierzalne uprzedzenia decyzyjne w zadaniach społecznych. <sup>18</sup>

W praktyce oznacza to, że „test intuicji” powinien kontrolować, czy model nie zmienia jakości argumentacji w zależności od sygnałów płciowych nadawcy/odbiorcy (to szczególnie ważne, jeśli dokumenty mają później działać jako dane dydaktyczne).

## **Implikacje projektowe: jak celowo kodować metadane w tekstach do meta-uczenia**

Twoje cztery artefakty (w szczególności „Wartość dodana...”) zawierają już załączek poprawnej praktyki: tekst ma być **przetwarzalny** w zadania, rubryki i kontrasty oraz ma zawierać „bezpieczniki epistemiczne” (jawne ograniczenia, falsyfikatory). (Dok. „Wartość dodana...”, s. 4–5; Dok. „Ocena formalności...”, s. 1)

Z perspektywy najlepszych praktyk w ML warto dodać dwa poziomy „metadanych nad tekstem” — nie jako metadanych pliku, lecz jako metadanych datasetu i próbki:

- **datasheets for datasets:** standard dokumentowania zbiorów danych (motywacja, skład, proces tworzenia, rekomendowane użycia, ryzyka), zaproponowany w literaturze jako środek transparentności i kontroli. <sup>19</sup>
- **model cards:** analogiczna praktyka dla modeli (zastosowania, ograniczenia, ewaluacja na grupach), opisywana jako narzędzie odpowiedzialnego raportowania. <sup>20</sup>

W Twoim kontekście „metadane w logice tekstu” (np. rubryka dowodu formalnego, legenda komunikacyjna, falsyfikatory) są tym, co model widzi w tokenach. Natomiast datasheet i model card to metadane, które zapewniają, że te tokeny nie zostaną użyte w sposób niezamierzony (np. metafora potraktowana jako fakt, albo adresatywność „Paniom” wzmocni niepożądane różnicowanie odpowiedzi).

Jako praktyczną „normę redakcyjną” dla tekstów człowiek+AI do meta-uczenia można przyjąć:

- **Jawne etykiety epistemiczne:** „to jest metafora”, „to jest hipoteza”, „to jest wniosek”, „to jest falsyfikator”. (Masz to miejscami już wprost; warto ujednolicić.) (Dok. „Paradoks Księżniczki...”, s. 6; Dok. „Paradoks Marii...”, s. 2)
- **Jedna rubryka oceny na dokument:** tak jak w „Ocenie formalności...”, bo to pozwala generować zadania testowe i krytykę o wysokiej powtarzalności. (Dok. „Ocena formalności...”, s. 1)
- **Kontrola biasu w metadanych adresatywnych:** jeśli „Paniom...” ma być sygnałem retorycznym, warto rozważyć neutralne warianty (np. „Szczególnie polecam antropologię...”) w danych treningowych, a wersję adresatywną zostawić jako osobny przypadek testowy do badania wrażliwości na płęć. Uzasadnieniem są wyniki literatury o gender bias i o tym, że redukcja biasu bywa nieintuicyjna (czasem „debiasing” może nawet pogorszyć sytuację downstream, jeśli osłabi szum maskujący sygnał płci). <sup>21</sup>

W skrócie: Twoje zdanie testowe jest miniaturą tego, co widzę w Twoich dokumentach — metadane sterują trybem czytania (antropologia/intuicja/test), a „logika artefaktu” daje się uporządkować jako system jawnych rubryk + ukrytych wzorców dyskursu. Antropologia jest tu dobrą metaforą, bo klasycznie uczy odróżniania „twitch” od „wink” — czyli powierzchni od kodu — a to dokładnie jest rdzeń metadanych tekstowych, które chcesz modelowi przekazać i następnie testować. <sup>22</sup>

---

<sup>1</sup> Frame analysis | Social Theory & Interactionism | Britannica  
[https://www.britannica.com/topic/frame-analysis?utm\\_source=chatgpt.com](https://www.britannica.com/topic/frame-analysis?utm_source=chatgpt.com)

<sup>2</sup> Zmiany w sposobach zwracania się do adresata na przykładzie mediów społecznościowych | Półrocznik Językoznawczy Tertium  
[https://journal.tertium.edu.pl/JaK/article/view/279?utm\\_source=chatgpt.com](https://journal.tertium.edu.pl/JaK/article/view/279?utm_source=chatgpt.com)

<sup>3</sup> Writing Without Conviction? Hedging in Science Research Articles | Applied Linguistics | Oxford Academic  
[https://academic.oup.com/applij/article-abstract/17/4/433/198756?utm\\_source=chatgpt.com](https://academic.oup.com/applij/article-abstract/17/4/433/198756?utm_source=chatgpt.com)

<sup>4</sup> <sup>5</sup> <sup>15</sup> <sup>22</sup> ICANN | Archives | Clifford Geertz: Thick Description: Toward an Interpretive Theory of Culture  
[https://archive.icann.org/en/ombudsman/clifford-geertz.htm?utm\\_source=chatgpt.com](https://archive.icann.org/en/ombudsman/clifford-geertz.htm?utm_source=chatgpt.com)

<sup>6</sup> <sup>7</sup> E-mentor - Etnografia w badaniu wiedzy ukrytej - E-mentor | nauczanie przez internet, kształcenie ustawiczne, metody, formy i programy  
[https://www.e-mentor.edu.pl/artukul/index/numer/43/id/906?utm\\_source=chatgpt.com](https://www.e-mentor.edu.pl/artukul/index/numer/43/id/906?utm_source=chatgpt.com)

- 8 Michael Polanyi | Philopedia  
[https://philopedia.org/thinkers/michael-polanyi/?utm\\_source=chatgpt.com](https://philopedia.org/thinkers/michael-polanyi/?utm_source=chatgpt.com)
- 9 Le Sens pratique  
[https://fr.wikipedia.org/wiki/Le\\_Sens\\_pratique?utm\\_source=chatgpt.com](https://fr.wikipedia.org/wiki/Le_Sens_pratique?utm_source=chatgpt.com)
- 10 Paper page - Attention Is All You Need  
[https://huggingface.co/papers/1706.03762?utm\\_source=chatgpt.com](https://huggingface.co/papers/1706.03762?utm_source=chatgpt.com)
- 11 NeurIPS Poster Training language models to follow instructions with human feedback  
[https://neurips.cc/virtual/2022/poster/52886?utm\\_source=chatgpt.com](https://neurips.cc/virtual/2022/poster/52886?utm_source=chatgpt.com)
- 12 OWASP Top 10 for Large Language Model Applications | OWASP Foundation  
[https://owasp.org/www-project-top-10-for-large-language-model-applications/?utm\\_source=chatgpt.com](https://owasp.org/www-project-top-10-for-large-language-model-applications/?utm_source=chatgpt.com)
- 13 UK cyber agency warns LLMs will always be vulnerable to prompt injection | CyberScoop  
[https://cyberscoop.com/uk-warns-ai-prompt-injection-unfixable-security-flaw/?utm\\_source=chatgpt.com](https://cyberscoop.com/uk-warns-ai-prompt-injection-unfixable-security-flaw/?utm_source=chatgpt.com)
- 14 PromptGuard a structured framework for injection resilient language models | Scientific Reports  
[https://www.nature.com/articles/s41598-025-31086-y?utm\\_source=chatgpt.com](https://www.nature.com/articles/s41598-025-31086-y?utm_source=chatgpt.com)
- 16 Self-Consistency Improves Chain of Thought Reasoning in Language Models - ADS  
[https://ui.adsabs.harvard.edu/abs/2022arXiv220311171W/abstract?utm\\_source=chatgpt.com](https://ui.adsabs.harvard.edu/abs/2022arXiv220311171W/abstract?utm_source=chatgpt.com)
- 17 AI models collapse when trained on recursively generated data | Nature  
[https://www.nature.com/articles/s41586-024-07566-y?utm\\_source=chatgpt.com](https://www.nature.com/articles/s41586-024-07566-y?utm_source=chatgpt.com)
- 18 Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings - Microsoft Research  
[https://www.microsoft.com/en-us/research/publication/quantifying-reducing-stereotypes-word-embeddings/?utm\\_source=chatgpt.com](https://www.microsoft.com/en-us/research/publication/quantifying-reducing-stereotypes-word-embeddings/?utm_source=chatgpt.com)
- 19 Datasheets for Datasets - ADS  
[https://ui.adsabs.harvard.edu/abs/2018arXiv180309010G/abstract?utm\\_source=chatgpt.com](https://ui.adsabs.harvard.edu/abs/2018arXiv180309010G/abstract?utm_source=chatgpt.com)
- 20 Model Cards for Model Reporting  
[https://research.google/pubs/model-cards-for-model-reporting/?utm\\_source=chatgpt.com](https://research.google/pubs/model-cards-for-model-reporting/?utm_source=chatgpt.com)
- 21 Debiasing Embeddings for Reduced Gender Bias in Text Classification - ACL Anthology  
[https://aclanthology.org/W19-3810/?utm\\_source=chatgpt.com](https://aclanthology.org/W19-3810/?utm_source=chatgpt.com)