

Artur Strzelecki

Sztuczna inteligencja w biznesie

Praktyczny
przewodnik
po narzędziach



onepress

Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości lub fragmentu niniejszej publikacji w jakiejkolwiek postaci jest zabronione. Wykonywanie kopii metodą kserograficzną, fotograficzną, a także kopiowanie książki na nośniku filmowym, magnetycznym lub innym powoduje naruszenie praw autorskich niniejszej publikacji.

Wszystkie znaki występujące w tekście są zastrzeżonymi znakami firmowymi bądź towarowymi ich właścicieli.

Autor oraz wydawca dołożyli wszelkich starań, by zawarte w tej książce informacje były kompletne i rzetelne. Nie biorą jednak żadnej odpowiedzialności ani za ich wykorzystanie, ani za związane z tym ewentualne naruszenie praw patentowych lub autorskich. Autor oraz wydawca nie ponoszą również żadnej odpowiedzialności za ewentualne szkody wynikłe z wykorzystania informacji zawartych w książce.

Redaktor prowadzący: Małgorzata Kulik
Recenzja naukowa: dr hab. Michał Kucia

Projekt okładki: Studio Gravite/Olsztyn
Obarek, Pokoński, Pazdrijowski, Zaprucki

Grafika na okładce została wykorzystana za zgodą AdobeStock.com.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: onepress@onepress.pl
WWW: onepress.pl (księgarnia internetowa, katalog książek)

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
onepress.pl/user/opinie/naaibi
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

ISBN: 978-83-289-3986-8

Copyright © Artur Strzelecki 2026

Printed in Poland.

- [Kup książkę](#)
- [Poleć książkę](#)
- [Oceń książkę](#)

- [Księgarnia internetowa](#)
- [Lubię to! » Nasza społeczność](#)

Spis treści

Wstęp	13
ROZDZIAŁ 1. Wprowadzenie do AI	17
1.1. Dlaczego sztuczna inteligencja zmienia biznes?	17
1.1.1. AI jako infrastruktura XXI wieku	17
1.1.2. Nowa energia dla biznesu	17
1.1.3. Przewaga konkurencyjna	18
1.1.4. Automatyzacja i produktywność	18
1.2. Czym jest sztuczna inteligencja?	18
1.2.1. Geneza terminu <i>Artificial Intelligence</i>	18
1.2.2. Definicja funkcjonalna sztucznej inteligencji	19
1.2.3. Test Turinga jako kryterium odniesienia	19
1.2.4. Współczesne rozumienie AI	20
1.3. Historia i przełomowe momenty w rozwoju AI	21
1.3.1. Przed narodzinami AI: fundamenty teoretyczne (1940 – 1955)	21
1.3.2. Konferencja w Dartmouth i narodziny dyscypliny (1956)	21
1.3.3. Pierwsza „zima AI” i symboliczne odrodzenie (1974 – 1980)	21
1.3.4. Druga „zima AI” i powrót sieci neuronowych (1987 – 1993)	22
1.3.5. Przełomowe momenty współczesnej ery AI (1997 – 2026)	22
1.3.6. Analiza wybranych kamieni milowych	23
1.3.7. Perspektywa historyczna	23
1.4. Fundament technologiczny sztucznej inteligencji	24
1.4.1. Hierarchia pojęć: AI $\supset$ ML $\supset$ DL $\supset$ LLM	24
1.4.2. Uczenie maszynowe: trzy paradygmaty	25
1.4.3. Deep Learning: wielowarstwowe reprezentacje	26
1.4.4. Architektura Transformer	26
1.4.5. Implikacje dla biznesu	28
1.5. Jak działają duże modele językowe (LLM)?	28
1.5.1. Przewidywanie kolejnego tokenu	29
1.5.2. Token: podstawowa jednostka przetwarzania	29
1.5.3. Prompt: interfejs komunikacji z modelem	30
1.5.4. Kontekst i okno kontekstowe	31
1.5.5. Parametry i wagi: wewnętrzna reprezentacja wiedzy	31
1.5.6. Temperatura: kontrola kreatywności	32
1.5.7. <i>Mixture of Experts</i>	32
1.5.8. RAG: wzbogacanie modelu o zewnętrzną wiedzę	33
1.6. Proces trenowania modeli językowych	34
1.6.1. Zbieranie danych	34
1.6.2. Uczenie wstępne: pre-training	35
1.6.3. Fine-tuning: dostrajanie do zastosowań	36
1.6.4. RLHF: uczenie się z ludzkiej informacji zwrotnej	36
1.6.5. <i>Alignment i safety filtering</i>	37
1.6.6. Cykl życia modelu	38

1.7. Koszty i zasoby trenowania modeli	38
1.7.1. Składniki kosztów pre-treningu	39
1.7.2. Szacunki kosztów treningu wybranych modeli	39
1.7.3. Energia i ślad środowiskowy	40
1.7.4. Globalne rynki pracy anotacyjnej	41
1.7.5. Bariery wejścia i koncentracja rynku	41
1.7.6. Perspektywy rozwoju rynku	42
1.8. Producenci modeli językowych	42
1.8.1. Stany Zjednoczone	42
1.8.2. Chiny	44
1.8.3. Europa	45
1.8.4. Inni znaczący producenci	46
1.8.5. Modele open source vs komercyjne	46
1.8.6. Rankingi i benchmarki	47
1.9. Typy sztucznej inteligencji w kontekście biznesowym	47
1.9.1. AI predykcyjna (<i>Predictive AI</i>)	48
1.9.2. AI generatywna (<i>Generative AI</i>)	49
1.9.3. AI agentowa (<i>Agentic AI</i>)	50
1.9.4. AI ucieleśniona (<i>Embodied AI</i>)	51
1.9.5. Porównanie typów AI	52
1.9.6. Konwergencja typów	52
1.10. Podsumowanie	52

ROZDZIAŁ 2. Inżynieria promptów 54

2.1. Wprowadzenie do inżynierii promptów	54
2.1.1. Definicja i geneza dyscypliny	54
2.1.2. Mechanizm działania	54
2.1.3. Znaczenie dla biznesu i organizacji	55
2.1.4. Podział technik na manualne i zautomatyzowane	56
2.2. Klasyfikacja metod inżynierii promptów	56
2.2.1. Metody manualne	56
2.2.2. Metody zautomatyzowane	57
2.3. Podstawowe techniki promptingu	58
2.3.1. Zero-shot prompting	58
2.3.2. One-shot prompting	59
2.3.3. Few-shot prompting	59
2.3.4. Porównanie wydajności podstawowych metod	60
2.4. Zaawansowane architektury rozumowania	61
2.4.1. <i>Chain-of-Thought</i> (CoT) Prompting	61
2.4.2. <i>Self-Consistency</i>	62
2.4.3. <i>Tree of Thoughts</i> (ToT)	62
2.4.4. <i>Prompt Chaining</i>	63
2.4.5. <i>Meta-prompting</i>	64
2.4.6. <i>ReAct</i> (<i>Reasoning and Acting</i>)	64
2.4.7. Porównanie architektur rozumowania	65
2.5. Frameworki strukturalne promptów	65
2.5.1. RTF (<i>Role, Task, Format</i>)	66
2.5.2. COSTAR (<i>Context, Objective, Style, Tone, Audience, Response</i>)	66
2.5.3. CARE (<i>Context, Action, Result, Example</i>)	67
2.5.4. TAG (<i>Task, Action, Goal</i>)	67
2.5.5. BAB (<i>Before, After, Bridge</i>)	68
2.5.6. RISE (<i>Role, Input, Steps, Expectation</i>)	68
2.5.7. Porównanie frameworków strukturalnych	69

2.6. Zasady i najlepsze praktyki projektowania	69
2.6.1. Fundamentalne zasady projektowania	69
2.6.2. Techniki strukturyzacji	70
2.6.3. Proces projektowania	70
2.6.4. Wykorzystanie bibliotek i narzędzi wspomagających	71
2.6.5. Unikanie stronniczości i aspekty etyczne	71
2.7. Automatyzacja i narzędzia	71
2.8. Ograniczenia, ryzyka i perspektywy	73
2.8.1. Fundamentalne ograniczenia promptingu	73
2.8.2. Bezpieczeństwo	74
2.8.3. PromptOps	74
2.8.4. Konwergencja typów AI i przyszłość inżynierii promptów	75
2.8.5. Rekomendacje strategiczne dla organizacji	76
2.9. Podsumowanie	76
ROZDZIAŁ 3. Generatywne przetwarzanie tekstu	77
3.1. Wprowadzenie do przetwarzania tekstu w biznesie	77
3.1.1. Rozwój od systemów regułowych do LLM	77
3.1.2. Przetwarzanie języka naturalnego w transformacji cyfrowej organizacji	77
3.1.3. Spektrum zastosowań	78
3.2. Mechanizmy przetwarzania	79
3.2.1. Token jako podstawowa jednostka przetwarzania	79
3.2.2. Okno kontekstowe	79
3.2.3. Małe vs duże okna kontekstowe	80
3.3. Architektura wdrożeniowa	81
3.3.1. Modele chmurowe	81
3.3.2. Modele lokalne	82
3.3.3. Porównanie modeli open source	85
3.3.4. Kwantyzacja jako technika optymalizacji	88
3.3.5. Kryteria wyboru architektury w sektorach wrażliwych	89
3.4. Środowiska testowe i porównawcze	90
3.4.1. Playground w cyklu życia modelu	90
3.4.2. Crowdsourcingowa ewaluacja	90
3.4.3. Narzędzia deweloperskie i komercyjne	91
3.4.4. Parametry sterujące generacją	91
3.5. Embeddingi i RAG	93
3.5.1. Modele embeddingowe	93
3.5.2. Ograniczenia tradycyjnej metody słów kluczowych	94
3.5.3. Architektura RAG	94
3.5.4. Modele E5 i Nomic	95
3.5.5. Zastosowania biznesowe RAG	96
3.6. Edytory tekstu wspomagane AI	96
3.6.1. Funkcjonalność współczesnych edytorów AI	96
3.6.2. Analiza semantyczna SERP	97
3.6.3. Proces współtworzenia treści	97
3.6.4. Integracja z modelami chmurowymi w edycji tekstu	98
3.7. Systemy przetwarzania dokumentów i zarządzania wiedzą	98
3.7.1. Proces przetwarzania dokumentów	98
3.7.2. Narzędzia interaktywne	99
3.7.3. Gemini Notebook jako zaawansowany system RAG	99
3.7.4. Mapa myśli i wizualizacja wiedzy	100

3.8. Podsumowanie	100
3.8.1. Integracja narzędzi przetwarzania tekstu w architekturę IT przedsiębiorstwa	100
3.8.2. Ochrona danych i suwerenność informacyjna	101
3.8.3. Perspektywa na przyszłość	101
ROZDZIAŁ 4. Generatywne przetwarzanie obrazu	102
4.1. Komunikacja wizualna	102
4.1.1. Rozwój technologii generatywnych	102
4.1.2. Przewagi biznesowe wdrożenia AI w produkcji wizualnej	102
4.1.3. Obszary aplikacji w środowisku korporacyjnym	103
4.2. Modele dyfuzyjne	103
4.2.1. Teoria procesu dyfuzji	104
4.2.2. Architektura U-Net	105
4.2.3. CLIP	105
4.2.4. ControlNet	105
4.2.5. Przewaga modeli dyfuzyjnych nad GAN	106
4.3. Inżynieria promptów graficznych	107
4.3.1. Hierarchia informacyjna	107
4.3.2. Specyfika składniowa narzędzi generatywnych	109
4.3.3. Język fotografii jako instrument kontroli	111
4.3.4. Techniki zaawansowane	112
4.3.5. Inżynieria odwrotna i automatyzacja optymalizacji	113
4.3.6. Proces iteracyjny od definicji do wariacji	113
4.4. Narzędzia chmurowe do generowania obrazów	115
4.4.1. Midjourney	115
4.4.2. OpenAI GPT Image	116
4.4.3. Modele open source w chmurze	116
4.4.4. Analiza porównawcza narzędzi chmurowych	118
4.5. Infrastruktura lokalna	119
4.5.1. Wymagania sprzętowe i konfiguracja środowiska	119
4.5.2. Interfejs użytkownika	120
4.5.3. Modele dostępne do wdrożeń lokalnych	120
4.5.4. LoRA i fine-tuning	121
4.5.5. Analiza opłacalności wdrożenia	121
4.6. Zastosowania biznesowe i wdrożenia praktyczne	122
4.6.1. E-commerce i fotografia produktowa	123
4.6.2. Marketing cyfrowy i branding	124
4.6.3. Komunikacja korporacyjna, szkolenia i edukacja	124
4.6.4. Zarządzanie spójnością tożsamości wizualnej	125
4.6.5. Framework MADE	126
4.6.6. Analiza efektywności i kluczowe wskaźniki sukcesu	127
4.7. Wyzwania techniczne, etyczne i prawne	128
4.7.1. Ograniczenia techniczne i artefakty generacji	128
4.7.2. Aspekty etyczne i prawne	131
4.7.3. Priorytety w procesie biznesowym	133
4.8. Podsumowanie	133
4.8.1. Evolucja paradygmatu	134
4.8.2. Konwergencja multimodalna i przyszłość interfejsów	134
4.8.3. Rekomendacje strategiczne dla przedsiębiorstw	135

ROZDZIAŁ 5. Generatywne przetwarzanie wideo	137
5.1. Wideo w komunikacji biznesowej	137
5.1.1. Znaczenie wideo w strategii marketingowej i edukacyjnej	137
5.1.2. Główne obszary aplikacji	138
5.1.3. Specyfika generatywna: temporalność i wyzwanie ciągłości	139
5.2. Podstawy technologiczne generowania wideo	139
5.2.1. Ewolucja dyfuzji 2D do modeli przestrzenno-czasowych	140
5.2.2. Specyfika procesu dyfuzji w wymiarze czasowym	140
5.2.3. Architektury 3D U-Net i rozszerzenia czasowe	141
5.2.4. Mechanizmy kontroli kompozycji w wideo	142
5.3. Inżynieria promptów wideo	142
5.3.1. Hierarchia informacyjna w prompcie wideo	142
5.3.2. Język kinematografii jako interfejs sterowania	143
5.3.3. System 10 filarów i frameworki strukturalne	144
5.3.4. Techniki zaawansowane	145
5.3.5. Meta-prompting i orkiestracja agentowa	146
5.3.6. <i>Negative prompting</i> i diagnostyka artefaktów	146
5.4. Narzędzia chmurowe do generowania wideo	147
5.4.1. Modele komercyjne	147
5.4.2. Analiza porównawcza narzędzi chmurowych	149
5.4.3. Wybór narzędzia w kontekście biznesowym	150
5.5. Infrastruktura lokalna i modele open source	150
5.5.1. Wymagania sprzętowe i konfiguracja środowiska	150
5.5.2. Modele dostępne do wdrożeń lokalnych	151
5.5.3. Interfejs użytkownika	152
5.5.4. Analiza opłacalności wdrożenia	153
5.6. Zastosowania biznesowe i wdrożenia praktyczne	154
5.6.1. Marketing cyfrowy i reklama performance	154
5.6.2. E-commerce i prezentacje produktowe	155
5.6.3. Edukacja korporacyjna i onboarding	156
5.6.4. Framework MADE w produkcji wideo	156
5.6.5. Studia przypadków: optymalizacja kampanii i lokalizacja	157
5.7. Wyzwania techniczne, etyczne i prawne	158
5.7.1. Ograniczenia techniczne i artefakty generacji	158
5.7.2. Aspekty etyczne i prawne	159
5.7.3. Priorytety w procesie biznesowym	160
5.8. Podsumowanie	161
5.8.1. Ewolucja paradygmatu	161
5.8.2. Konwergencja multimodalna i przyszłość interfejsów	162
5.8.3. Rekomendacje strategiczne dla przedsiębiorstw	163
ROZDZIAŁ 6. Generatywne przetwarzanie audio i synteza mowy	164
6.1. Audio i głos w kontekście biznesowym	164
6.1.1. Jednolity workflow audio-głos	164
6.1.2. <i>Text-to-Audio vs Text-to-Speech</i>	165
6.1.3. Obszary aplikacji	166
6.2. Podstawy technologiczne syntezy audio i mowy	167
6.2.1. Ewolucja generatywnych modeli audio	167
6.2.2. Architektury generatywne	168
6.2.3. Warstwy akustyczne	169
6.3. Inżynieria promptów audio (<i>Text-to-Audio</i>)	170
6.3.1. Filary strukturalne promptu muzycznego	171
6.3.2. Metatagi i instrukcje strukturalne	172
6.3.3. Model S-A-D-E dla efektów dźwiękowych	173

6.3.4. Parametryzacja rytmiczna i lingwistyczna	174
6.3.5. <i>Negative prompting</i> w muzyce	175
6.4. Inżynieria promptów głosowych (<i>Text-to-Speech</i> i <i>Speech-to-Speech</i>)	176
6.4.1. Deskryptywne projektowanie głosu	176
6.4.2. Język znaczników SSML (<i>Speech Synthesis Markup Language</i>)	178
6.4.3. Promptowanie in-line	179
6.4.4. Metoda <i>Speech-to-Speech</i> (STS)	180
6.4.5. Zarządzanie wymową	181
6.5. Narzędzia chmurowe do generowania audio i głosu	183
6.5.1. Głos	183
6.5.2. Muzyka i efekty	185
6.5.3. Analiza porównawcza	186
6.6. Infrastruktura lokalna	188
6.6.1. Modele open source	188
6.6.2. Wymagania sprzętowe i konfiguracja środowiska	189
6.6.3. Prywatność danych	190
6.7. Zastosowania biznesowe i wdrożenia praktyczne	192
6.7.1. Produkcja reklamowa	192
6.7.2. Obsługa klienta i IVR	193
6.7.3. Edukacja i szkolenia	194
6.7.4. Dubbing i lokalizacja	195
6.7.5. Inpainting audio	196
6.8. Wyzwania techniczne, etyczne i prawne	197
6.8.1. <i>Deepfake voice</i> i tożsamość głosowa	197
6.8.2. Prawa autorskie do głosu i zbiorów treningowych	198
6.8.3. Ograniczenia techniczne	199
6.8.4. Compliance i transparentność	200
6.9. Podsumowanie	201
6.9.1. Konwergencja audio-wideo-głos	202
6.9.2. Przyszłość	202
6.9.3. Rekomendacje dla biznesu	203

ROZDZIAŁ 7. Asystenty i agenty AI205

7.1. Wprowadzenie	205
7.1.1. Trzy poziomy pracy z AI	205
7.1.2. Oś dojrzałości pracy z AI	206
7.1.3. Czego oczekuje dziś firma od AI?	207
7.2. Asystenty AI — definicja i konfiguracja	207
7.2.1. Definicja asystenta AI	207
7.2.2. Konfigurowalne asystenty i personalizacja	208
7.2.3. Zastosowania biznesowe i ograniczenia asystentów	209
7.3. Systemy Deep Research	210
7.3.1. Definicja Deep Research	210
7.3.2. Cztery filary techniczne Deep Research	211
7.3.3. Architektura i workflow systemów Deep Research	212
7.3.4. Zastosowania biznesowe systemów Deep Research	213
7.3.5. Kryteria oceny jakości Deep Research	213
7.3.6. Ryzyka i ograniczenia systemów Deep Research	214
7.3.7. Przykłady systemów Deep Research na rynku	215
7.4. Agenty AI — definicja i architektura	215
7.4.1. Definicja agenta AI	215
7.4.2. Poziomy autonomii agentów AI	216
7.4.3. Architektura agenta AI	217
7.4.4. Workflow działania agenta	218

7.4.5. Zastosowania biznesowe i korzyści z wdrożenia agentów	218
7.4.6. Ryzyka i ograniczenia architektury agentowej	219
7.5. OpenClaw — autonomiczny agent osobisty	219
7.5.1. Geneza i filozofia systemu	219
7.5.2. Architektura techniczna — trzy filary	220
7.5.3. Autonomia operacyjna i scenariusze biznesowe w MŚP	221
7.5.4. Wyzwania bezpieczeństwa i ryzyka wdrożeniowe	222
7.5.5. Aspekty ekonomiczne i infrastrukturalne wdrożenia	223
7.6. Model Context Protocol (MCP) — standard komunikacji	224
7.6.1. Definicja i pochodzenie standardu	224
7.6.2. Architektura klient-serwer	225
7.6.3. Funkcjonalne możliwości protokołu	225
7.6.4. Znaczenie dla standaryzacji agentów AI	226
7.6.5. Praktyczne zastosowania i przykłady wdrożenia	226
7.6.6. Implikacje strategiczne dla przedsiębiorstw	227
7.7. Orkiestracja agentów — systemy wieloagentowe	227
7.7.1. Ograniczenia pojedynczego agenta	228
7.7.2. Definicja i zakres orkiestracji agentów	228
7.7.3. Typowe role w systemie wieloagentowym	229
7.7.4. Wzorce architektury systemów wieloagentowych	230
7.7.5. <i>Agentic engineering</i> — projektowanie systemów agentowych	231
7.7.6. Ryzyka i wyzwania orkiestracji wieloagentowej	232
7.8. Podsumowanie	232
7.8.1. Synteza kluczowych koncepcji	233
7.8.2. Implikacje strategiczne dla przedsiębiorstw	233
7.8.3. Kierunki dalszego rozwoju	234
ROZDZIAŁ 8. <i>Low-code, no-code i vibe coding</i>	235
8.1. Wprowadzenie do użytkowania i tworzenia	235
8.2. <i>Low-code i no-code</i> — definicje i różnice	236
8.2.1. <i>Low-code</i> jako czwarta generacja języków programowania	236
8.2.2. <i>No-code</i> jako całkowita eliminacja bariery kodowania	236
8.2.3. Kluczowe rozróżnienia techniczne i organizacyjne	237
8.2.4. Persony użytkowników i demokratyzacja rozwoju	238
8.3. Zalety i ograniczenia platform <i>low-code</i> oraz <i>no-code</i>	238
8.3.1. Zalety biznesowe platform <i>low-code</i> i <i>no-code</i>	238
8.3.2. Ograniczenia i wyzwania wdrożeniowe	239
8.4. Platformy <i>low-code</i> i <i>no-code</i> w otoczeniu AI	241
8.4.1. Make — wizualna automatyzacja workflow z elementami agentowymi	241
8.4.2. Dify — platforma do budowy aplikacji AI	242
8.4.3. n8n — automatyzacja workflow dla zespołów technicznych	242
8.4.4. Kryteria wyboru platformy	243
8.5. <i>Vibe coding</i> — programowanie konwersacyjne	244
8.5.1. Definicja i geneza paradygmatu	244
8.5.2. Fazy rozwoju: Day 0 i Day 1+	245
8.5.3. Narzędzia infrastruktury <i>vibe coding</i>	245
8.5.4. <i>Vibe coding</i> a <i>low-code</i> i <i>no-code</i>	246
8.5.5. Ryzyka i ograniczenia <i>vibe coding</i>	246
8.5.6. Studia przypadków awarii produkcyjnych	247
8.6. Zastosowania biznesowe i wybór strategii	250
8.6.1. Mapowanie potrzeb biznesowych na paradygmaty technologiczne	250
8.6.2. Scenariusze wdrożeniowe i kryteria decyzyjne	251

8.6.3. Hybrydyzacja strategii i kontinuum technologiczne	251
8.6.4. Implikacje dla zarządzania IT	252
8.7. Podsumowanie	252
8.7.1. Synteza kluczowych koncepcji	252
8.7.2. Kompromisy strategiczne	253
8.7.3. Implikacje dla zarządzania	253
ROZDZIAŁ 9. Kreatory AI	255
9.1. Od kodowania do konwersacji	255
9.1.1. Przesunięcie paradygmatu od programowania do inżynierii podpowiedzi	255
9.1.2. Demokratyzacja tworzenia oprogramowania	256
9.2. Architektura i modele wykorzystywane przez kreatory AI	257
9.2.1. Mechanizm wyboru modelu	259
9.2.2. Logowanie przez Google jako standard dostępu	259
9.3. Porównanie kreatorów AI	260
9.3.1. Zestawienie porównawcze	261
9.3.2. Analiza różnic funkcjonalnych	263
9.4. Kreatory <i>full-stack</i>	264
9.4.1. Replit	264
9.4.2. Emergent	265
9.4.3. Floom	265
9.4.4. Base44	266
9.4.5. Bolt	266
9.4.6. Mocha	267
9.4.7. Zestawienie porównawcze	267
9.4.8. Wnioski dla przedsiębiorstw	268
9.5. Kreatory frontendowe i projektowe	268
9.5.1. v0 (Vercel)	269
9.5.2. Magic Patterns	269
9.5.3. Figma AI	270
9.5.4. Canva	270
9.5.5. Zestawienie porównawcze	271
9.5.6. Wnioski dla przedsiębiorstw	271
9.6. Kreatory stron i treści wizualnych	272
9.6.1. Loveable	272
9.6.2. Aura	273
9.6.3. Macaly	273
9.6.4. Canva	274
9.6.5. Zestawienie porównawcze	274
9.6.6. Wnioski dla przedsiębiorstw	275
9.7. Modele autorskie i systemy otwarte	275
9.7.1. Canva	276
9.7.2. Cosmic	276
9.7.3. Zestawienie porównawcze	277
9.7.4. Wnioski dla przedsiębiorstw	278
9.8. Claude Design	278
9.8.1. Funkcjonalności i mechanizmy produkcyjne	278
9.8.2. Reakcja rynku	279
9.8.3. Segment docelowy i implikacje dla struktury rynku	279
9.8.4. Wnioski dla przedsiębiorstw	280
9.9. Wnioski i rekomendacje biznesowe	280
9.9.1. Kryteria wyboru kreatora AI	281
9.9.2. Ryzyka i ograniczenia	282

9.9.3. Perspektywy rozwoju i scenariusze ewolucji rynku	282
9.9.4. Rekomendacje końcowe	283
ROZDZIAŁ 10. Benchmarki AI	284
10.1. Wprowadzenie	284
10.1.1. Zjawisko saturacji i cykl życia benchmarku	284
10.1.2. Dlaczego benchmarki są kluczowe dla decyzji biznesowych?	285
10.1.3. Raporty i indeksy	285
10.1.4. Pułapki i ograniczenia wstępne	286
10.2. Kod i agenty	286
10.2.1. Zestawienie benchmarków programistycznych	287
10.2.2. Analiza szczegółowa wybranych benchmarków	287
10.2.3. Znaczenie dla biznesu i oceny agentów programistycznych	290
10.3. Ogólne agenty	291
10.3.1. Zestawienie benchmarków ogólnoagentowych	291
10.3.2. Analiza szczegółowa wybranych benchmarków	292
10.3.3. Znaczenie dla biznesu	295
10.4. Wiedza i QA	296
10.4.1. Zestawienie benchmarków wiedzy i QA	296
10.4.2. Analiza szczegółowa wybranych benchmarków	297
10.4.3. Znaczenie dla biznesu	298
10.5. Instrukcje i długi kontekst	299
10.5.1. Zestawienie benchmarków instrukcyjnych i długiego kontekstu	299
10.5.2. Analiza szczegółowa	300
10.5.3. Znaczenie dla biznesu	301
10.6. STEM i rozumowanie	302
10.6.1. Zestawienie benchmarków STEM i rozumowania	303
10.6.2. Analiza szczegółowa wybranych benchmarków	303
10.6.3. Znaczenie dla biznesu	306
10.7. Wielojęzyczność	307
10.7.1. Zestawienie benchmarków wielojęzycznych	307
10.7.2. Analiza szczegółowa wybranych benchmarków	308
10.7.3. Znaczenie dla biznesu	310
10.8. Zjawisko zawyżania wyników i systemowa krytyka benchmarków	310
10.8.1. Mechanizmy zawyżania wyników	311
10.8.2. Kontaminacja danych i utrata zdolności generalizacji	311
10.8.3. Selektywne raportowanie i koluzja	312
10.8.4. <i>Sandbagging</i> i strategiczna manipulacja ewaluacją	312
10.8.5. Wady systemowe	313
10.8.6. Wnioski dla przedsiębiorstw i regulatorów	314
10.9. Globalne indeksy i raporty AI	314
10.9.1. Stanford AI Index 2026	314
10.9.2. Artificial Analysis Intelligence Index 4.0	315
10.9.3. Wnioski dla biznesu	316
10.10. Seria ARC-AGI	317
10.10.1. RHAЕ i efektywność działania jako miara inteligencji	317
10.10.2. Znaczenie dla biznesu	318
10.11. Statystyczne frameworki agregujące	318
10.11.1. Problem fragmentacji i saturacji	318
10.11.2. Model statystyczny	319
10.11.3. Zastosowania frameworku agregującego	319
10.11.4. Ograniczenia i krytyka metodologiczna	321
10.11.5. Znaczenie dla biznesu i regulacji	321

10.12. Przyszłość benchmarków (PeerBench)	322
10.12.1. Krytyka obecnych leaderboardów jako „otwartych ksiąg”	322
10.12.2. PeerBench — platforma oparta na zasadach egzaminów humanistycznych	323
10.12.3. Wyzwania implementacyjne i perspektywy	324
10.13. Wnioski i rekomendacje dla biznesu	324
10.13.1. Sześć zasad rzetelnej oceny modeli w przedsiębiorstwie	324
10.13.2. Perspektywy regulacyjne i strategiczne	325
10.13.3. Rekomendacje końcowe	326
ROZDZIAŁ 11. AI jako siła strategiczna organizacji	327
11.1. Tempo i skala transformacji	327
11.1.1. AI jako technologia ogólnego zastosowania w gospodarce	327
11.1.2. Przyspieszenie adopcji	328
11.1.3. Wymiar ekonomiczny	329
11.2. Architektura władzy w ekosystemie AI	330
11.2.1. Cztery filary przewagi w AI	330
11.2.2. Bigtechy i modele podstawowe	330
11.2.3. Łańcuch wartości AI	331
11.2.4. Suwerenność AI	331
11.3. Regulacje w świecie niedostosowanym do tempa technologii	332
11.3.1. Dlaczego AI wymaga osobnych ram regulacyjnych?	333
11.3.2. Logika AI Act	333
11.3.3. Obowiązki organizacji	335
11.3.4. Piaskownice regulacyjne	336
11.3.5. Regulatory <i>backsliding</i>	337
11.4. Kompetencje i przebudowa pracy	338
11.4.1. Luka przygotowania organizacyjnego	338
11.4.2. Od dostępu do aktywacji	339
11.4.3. Trzy poziomy transformacji organizacyjnej	340
11.4.4. AI a rynek pracy	342
11.5. Europa w globalnym układzie sił	344
11.5.1. AI Continent Action Plan	344
11.5.2. Europa na mapie globalnej adopcji	345
11.5.3. Suwerenna infrastruktura AI	346
11.5.4. Minimalny program zgodności AI Act dla przedsiębiorstwa	348
11.6. Nowa fala AI	349
11.6.1. Agenty AI	350
11.6.2. Physical AI	351
11.6.3. Wyzwania governance przy autonomii maszyn	352
11.7. Podsumowanie	353
Bibliografia	355

Wstęp

Z przyjemnością oddaję w Twoje ręce książkę *Sztuczna inteligencja w biznesie. Praktyczny przewodnik po narzędziach AI*. Publikacja ta powstała z potrzeby uporządkowania wiedzy o technologii, która w niezwykle krótkim czasie stała się jednym z najważniejszych zjawisk współczesnej gospodarki cyfrowej. Sztuczna inteligencja to dzisiaj codzienne narzędzie pracy, element infrastruktury przedsiębiorstw oraz siła, która wpływa na sposób tworzenia treści, przetwarzania informacji, podejmowania decyzji, projektowania produktów i organizowania procesów biznesowych.

O sztucznej inteligencji mówi się dziś bardzo dużo. Jest ona obecna w mediach, debatach publicznych, strategiach przedsiębiorstw, programach edukacyjnych i codziennych rozmowach. Popularność tego tematu nie wynika jednak wyłącznie z chwilowej mody. Za zainteresowaniem AI stoją realne zmiany technologiczne i gospodarcze. Systemy oparte na dużych modelach językowych potrafią przetwarzać tekst, obrazy, dźwięk i wideo, analizować dokumenty, tworzyć oprogramowanie, prowadzić badania, komunikować się z innymi systemami oraz wykonywać coraz bardziej złożone zadania. W wielu organizacjach sztuczna inteligencja staje się nową warstwą dostępu do wiedzy, danych, narzędzi i procesów.

Umiejętność świadomego korzystania z AI staje się zatem jedną z kluczowych kompetencji współczesnego człowieka. Dotyczy to zarówno młodych ludzi, którzy dopiero przygotowują się do wejścia na rynek pracy, jak i menedżerów, przedsiębiorców, analityków, specjalistów, nauczycieli, badaczy i wszystkich osób uczestniczących w gospodarce cyfrowej. Znaczenie tej kompetencji obejmuje rozumienie mechanizmów działania modeli, umiejętność formułowania instrukcji, ocenę jakości rezultatów, dobór właściwych narzędzi, ochronę danych, identyfikowanie ryzyka oraz odpowiedzialne włączanie AI do procesów organizacyjnych.

Pomysł napisania tej książki zrodził się z obserwacji, jak szybko zmienia się środowisko narzędzi sztucznej inteligencji oraz jak trudno jest uporządkować informacje pochodzące z licznych, często rozproszonych źródeł. Każdego miesiąca pojawiają się nowe modele, platformy, metody pracy i zastosowania. Rozwiązania, które niedawno wydawały się eksperymentalne, stają się częścią powszechnie dostępnych usług, a granice między tworzeniem treści, analizą informacji, automatyzacją i programowaniem stopniowo się zacierają. Jednocześnie szybkie tempo rozwoju powoduje, że wiedza o AI bywa fragmentaryczna, uproszczona albo nadmiernie skoncentrowana na pojedynczych narzędziach.

Celem tej książki jest przedstawienie sztucznej inteligencji w sposób możliwie całościowy, a jednocześnie zrozumiały i praktyczny. Książka łączy podstawy technologiczne z praktyką biznesową, pokazując zarówno mechanizmy działania współczesnych systemów AI, jak i sposoby ich zastosowania w organizacji. Czytelnik poznaje nie tylko możliwości narzędzi, lecz także ich ograniczenia, koszty, ryzyka, wymagania infrastrukturalne

oraz konsekwencje organizacyjne i regulacyjne. Publikacja została podzielona na jedenaście rozdziałów, które tworzą uporządkowaną drogę od podstaw sztucznej inteligencji do jej strategicznego znaczenia dla przedsiębiorstw i gospodarki.

Pierwszy rozdział wprowadza czytelnika w świat sztucznej inteligencji. Przedstawia genezę tej dziedziny, najważniejsze etapy jej rozwoju oraz przełomowe wydarzenia, które doprowadziły do powstania współczesnych modeli. Wyjaśnia relacje pomiędzy sztuczną inteligencją, uczeniem maszynowym, uczeniem głębokim i dużymi modelami językowymi. Omawia również sposób działania modeli, znaczenie tokenów, kontekstu, parametrów, temperatury, architektury transformerów i mechanizmu *retrieval-augmented generation*. Rozdział uwzględnia także proces trenowania modeli, jego koszty ekonomiczne i środowiskowe, strukturę globalnego rynku oraz podział na AI predykcyjną, generatywną, agentową i ucieleśnioną.

Drugi rozdział został poświęcony inżynierii promptów, czyli sztuce i metodologii komunikowania się z modelami. Czytelnik poznaje podstawowe techniki promptowania, takie jak *zero-shot*, *one-shot* i *few-shot*, a następnie bardziej zaawansowane podejścia wspierające rozumowanie i wykonywanie złożonych zadań. Przedstawione zostały także strukturalne frameworki tworzenia promptów, zasady projektowania instrukcji, narzędzia automatyzujące ten proces oraz zagrożenia związane z bezpieczeństwem, stronniczością i nieprzewidywalnością wyników. Rozdział pokazuje, że skuteczna współpraca z AI wymaga odpowiedniego sformułowania pytania, a także rozumienia celu, kontekstu, ograniczeń i oczekiwanego formatu odpowiedzi.

Kolejne cztery rozdziały dotyczą generatywnego przetwarzania tekstu, obrazu, wideo oraz dźwięku i mowy. W części poświęconej tekstowi omówione są modele chmurowe i lokalne, kwantyzacja, środowiska testowe, embeddingi, systemy RAG, edytory wspierane przez AI oraz narzędzia do przetwarzania dokumentów i zarządzania wiedzą. Rozdział dotyczący obrazu przedstawia podstawy modeli dyfuzyjnych, zasady graficznej inżynierii promptów, porównanie narzędzi chmurowych i lokalnych oraz ich zastosowania w marketingu, e-commerce, komunikacji i budowaniu spójności wizualnej marki. Rozdział poświęcony wideo pokazuje, w jaki sposób modele generatywne radzą sobie z czasem, ruchem, ciągłością scen i kompozycją. Omawia język filmowy jako narzędzie sterowania generacją, porównuje najważniejsze klasy systemów oraz przedstawia zastosowania w reklamie, prezentacji produktów, szkoleniach i lokalizacji treści. Z kolei rozdział dotyczący audio i syntezy mowy obejmuje generowanie muzyki, efektów dźwiękowych, głosu i wypowiedzi. Czytelnik poznaje zasady projektowania promptów audio, możliwości syntezy i transformacji mowy, zastosowania w reklamie, obsłudze klienta, edukacji i dubbingu, a także problemy związane z deepfake'ami głosowymi, tożsamością i prawem do głosu.

Siódmy rozdział prowadzi od prostych asystentów do systemów zdolnych do samodzielnego prowadzenia badań i wykonywania działań. Omawia konfigurowalne asystenty, systemy deep research, architekturę agentów, poziomy ich autonomii, pamięć, planowanie, korzystanie z narzędzi i realizację wieloetapowych zadań. Przedstawia także Model Context Protocol jako standard komunikacji modeli z zewnętrznymi zasobami oraz systemy wieloagentowe, w których wiele wyspecjalizowanych agentów współpracuje przy realizacji wspólnego celu. Rozdział pozwala zrozumieć przejście od AI odpowiadającej na pytania do AI zdolnej do podejmowania działań.

Ósmy rozdział został poświęcony platformom *low-code* i *no-code* oraz zjawisku *vibe coding*. Pokazuje, w jaki sposób sztuczna inteligencja obniża barierę wejścia do tworzenia aplikacji i automatyzowania procesów. Porównuje różne podejścia do budowania rozwiązań cyfrowych, omawia ich korzyści, ograniczenia, ryzyka oraz konsekwencje dla zarządzania technologią w organizacji. Szczególna uwaga została zwrócona na różnice pomiędzy szybkim stworzeniem prototypu a utrzymaniem bezpiecznego i skalowalnego rozwiązania produkcyjnego.

Rozdział dziewiąty rozwija ten temat poprzez analizę AI builders, czyli platform umożliwiających tworzenie stron internetowych, interfejsów, aplikacji i kompletnych produktów cyfrowych za pomocą języka naturalnego. Poszczególne narzędzia zostały podzielone według ich przeznaczenia i możliwości, a następnie porównane pod względem funkcjonalności, zastosowanych modeli, integracji, własności kodu, skalowalności, kosztów i przydatności dla przedsiębiorstw. Rozdział pokazuje, że granica między użytkownikiem a twórcą oprogramowania staje się coraz mniej wyraźna.

Dziesiąty rozdział dotyczy benchmarków, rankingów i metod oceny systemów sztucznej inteligencji. Jest to szczególnie ważny obszar z punktu widzenia przedsiębiorstw, które muszą wybierać modele i narzędzia na podstawie bardziej wiarygodnych kryteriów niż popularność czy deklaracje producentów. W rozdziale przedstawiono benchmarki dotyczące programowania, agentów, wiedzy, wykonywania instrukcji, długiego kontekstu, rozumowania, matematyki, wielojęzyczności i przetwarzania multimodalnego. Omówiono także zjawiska inflacji wyników, zanieczyszczenia danych, selektywnego raportowania i strategicznego manipulowania oceną modeli. Celem tej części jest rozwinięcie umiejętności krytycznego odczytywania rankingów i samodzielnego oceniania przydatności AI.

Ostatni rozdział przedstawia sztuczną inteligencję jako siłę strategiczną organizacji. Analizuje tempo adopcji, gospodarcze znaczenie AI, strukturę globalnego zainteresowania AI, koncentrację rynku, rolę wielkich firm technologicznych oraz problem suwerenności technologicznej. Omawia regulacje, w tym logikę unijnego AI Act, obowiązki organizacji oraz znaczenie piaskownic regulacyjnych. Wiele miejsca poświęcono kompetencjom, zmianom na rynku pracy, przygotowaniu organizacji do wdrażania AI oraz różnicy między samym dostępem do technologii a jej rzeczywistym wykorzystaniem. Rozdział kończy się refleksją nad kolejną falą rozwoju, obejmującą agentów, fizyczną AI oraz systemy charakteryzujące się coraz większą autonomią.

Choć poszczególne rozdziały dotyczą różnych technologii i obszarów zastosowań, tworzą one spójną całość. Książka prowadzi od zrozumienia podstaw, poprzez nabywanie praktycznych umiejętności, aż do krytycznej oceny narzędzi i refleksji nad ich znaczeniem dla organizacji. Można ją czytać kolejno, traktując jako systematyczny podręcznik, ale poszczególne rozdziały mogą również służyć jako samodzielne przewodniki po konkretnych obszarach sztucznej inteligencji.

Mam nadzieję, że książka stanie się źródłem wiedzy o aktualnych narzędziach i przewodnikiem pozwalającym lepiej zrozumieć kierunek zachodzących zmian. Konkretnie platformy i modele będą się zmieniać, jednak kompetencje związane z formułowaniem problemów, oceną informacji, wyborem technologii, projektowaniem procesów oraz odpowiedzialnym korzystaniem z AI zachowają swoje znaczenie.

Jestem przekonany, że sztuczna inteligencja nie zastąpi potrzeby posiadania wiedzy, krytycznego myślenia i doświadczenia. Może jednak znacząco rozszerzyć możliwości osób,

które potrafią korzystać z niej świadomie. Dlatego mam nadzieję, że niniejsza publikacja pomoże czytelnikom nie tylko poznać współczesne narzędzia AI, lecz także lepiej przygotować się do pracy, nauki i podejmowania decyzji w świecie, w którym sztuczna inteligencja staje się coraz ważniejszą częścią gospodarki i życia społecznego.

Życzę inspirującej lektury,
Artur Strzelecki
profesor Uniwersytetu Ekonomicznego w Katowicach

Wprowadzenie do AI

1.1. Dlaczego sztuczna inteligencja zmienia biznes?

Sztuczna inteligencja (AI) wydostała się z obszaru badań akademickich i stała się fundamentalną infrastrukturą gospodarki XXI wieku. Podobnie jak elektryczność zrewolucjonizowała przemysł na przełomie XIX i XX wieku, a internet przekształcił gospodarkę pod koniec XX wieku, sztuczna inteligencja obecnie redefiniuje sposób funkcjonowania przedsiębiorstw, modeli biznesowych oraz relacji między organizacjami a ich klientami (Chalmers i in., 2026).

1.1.1. AI jako infrastruktura XXI wieku

Współczesne przedsiębiorstwa coraz częściej traktują sztuczną inteligencję nie jako opcjonalne narzędzie, lecz jako kluczowy element infrastruktury operacyjnej. AI umożliwia przetwarzanie i analizę danych w skali niemożliwej do osiągnięcia przez człowieka, automatyzację złożonych procesów decyzyjnych oraz tworzenie spersonalizowanych doświadczeń dla klientów. Zastosowania te obejmują niemal wszystkie sektory gospodarki — od finansów i medycyny, przez produkcję i logistykę, aż do kreatywnych przemysłów kulturalnych.

1.1.2. Nowa energia dla biznesu

Metafora „nowej energii” oddaje transformacyjny charakter sztucznej inteligencji dla przedsiębiorstw. AI dostarcza organizacjom zdolności do:

- **przetwarzania nieustrukturyzowanych danych** — tekstów, obrazów, dźwięków i wideo — w sposób umożliwiający ekstrakcję wartościowych informacji biznesowych;
- **generowania treści** — od dokumentacji technicznej i kodu programistycznego, przez materiały marketingowe, aż po projekty graficzne i kompozycje muzyczne;
- **prowadzenia interakcji w języku naturalnym** — podczas obsługi klienta, udzielania wsparcia decyzyjnego oraz w trakcie wewnętrznej komunikacji organizacyjnej.

Te możliwości przekładają się na konkretne korzyści ekonomiczne, jak redukcja kosztów operacyjnych, przyspieszenie cykli innowacyjnych oraz otwarcie nowych obszarów działalności biznesowej.

1.1.3. Przewaga konkurencyjna

W dynamicznym środowisku rynkowym wdrożenie narzędzi AI staje się czynnikiem decydującym o utrzymaniu przewagi konkurencyjnej. Przedsiębiorstwa wykorzystujące sztuczną inteligencję mogą:

- podejmować decyzje oparte na danych w czasie rzeczywistym;
- dostosowywać ofertę do indywidualnych potrzeb klientów z precyzją trudną do uzyskania przy tradycyjnych metodach segmentacji;
- prognozować trendy rynkowe i zachowania konkurentów z wyprzedzeniem umożliwiającym strategiczne działania.

Badania wskazują, że organizacje zaawansowane we wdrażaniu AI charakteryzują się wyższą produktywnością, szybszym tempem wzrostu oraz większą odpornością na zakłócenia rynkowe (Chalmers i in., 2026).

1.1.4. Automatyzacja i produktywność

Jednym z najbardziej wymiernych aspektów wpływu AI na biznes jest potencjał automatyzacyjny. Sztuczna inteligencja umożliwia:

- **automatyzację zadań poznawczych** — analizy dokumentów, klasyfikacji informacji, ekstrakcji danych;
- **wsparcie procesów decyzyjnych** — scoringu kredytowego, wykrywania oszustw, optymalizacji łańcuchów dostaw;
- **generowanie i kurację treści** — tworzenia wstępnych wersji dokumentów, które następnie weryfikowane są przez człowieka.

Istotne jest, że automatyzacja wspierana przez AI nie polega wyłącznie na zastępowaniu pracy ludzkiej, lecz często na jej **rozszerzeniu** — wzmocnieniu możliwości pracowników poprzez eliminację powtarzalnych zadań i uwolnienie kapitału ludzkiego dla działań wymagających kreatywności, empatii i złożonego rozumowania.

1.2. Czym jest sztuczna inteligencja?

Definicja sztucznej inteligencji ewoluowała na przestrzeni dekad, od pierwotnych koncepcji teoretycznych po współczesne systemy o złożonych możliwościach generatywnych i agentowych. Zrozumienie genezy terminu oraz zakresu zadań przypisywanych AI stanowi fundament dla dalszej analizy narzędzi i zastosowań biznesowych.

1.2.1. Geneza terminu *Artificial Intelligence*

Termin *Artificial Intelligence* (sztuczna inteligencja) został wprowadzony do obiegu naukowego przez **Johna McCarthy’ego** w 1956 roku podczas letniej konferencji w Dartmouth College (McCarthy i in., 2006). McCarthy, profesor matematyki, zaproponował organizację konferencji wspólnie z Marvinem Minskym (Harvard), Nathanielem Rochesterem

(IBM) oraz Claude’em Shannonem (Bell Labs) — postaciami kluczowymi dla rozwoju informatyki i teorii informacji.

Celem konferencji było zbadanie założenia, że „każdy aspekt uczenia się lub innej cechy inteligencji może w zasadzie być tak precyzyjnie opisany, że maszyna będzie w stanie go symulować” (McCarthy i in., 2006). Choć konferencja nie przyniosła przełomowych wyników natychmiastowych, ustanowiła paradygmat badawczy i terminologiczny, który definiuje pole do dziś.

Warto zauważyć, że wśród uczestników konferencji znalazł się **Claude Shannon** — matematyk, którego imię zostało później użyte przez firmę Anthropic jako nazwa modelu językowego *Claude*.

1.2.2. Definicja funkcjonalna sztucznej inteligencji

W ujęciu funkcjonalnym sztuczna inteligencja obejmuje **systemy wykonujące zadania wymagające ludzkiej inteligencji**. Definicja ta, choć szeroka, pozwala na operacjonalizację poprzez wskazanie czterech kluczowych obszarów zdolności (tabela 1.1).

TABELA 1.1. Cztery obszary funkcjonalności AI

OBSZAR ZDOLNOŚCI	OPIS	PRZYKŁADY ZASTOSOWAŃ BIZNESOWYCH
Rozumowanie	Zdolność do wnioskowania, analizy logicznej i rozwiązywania problemów	Analiza umów prawnych, wsparcie decyzyjne w zarządzaniu
Rozpoznawanie wzorców	Identyfikacja regularności w danych strukturalnych i nieustrukturyzowanych	Wykrywanie oszustw finansowych, segmentacja klientów
Podejmowanie decyzji	Wybór optymalnych działań na podstawie dostępnych informacji i celów	Zautomatyzowany handel, optymalizacja łańcuchów dostaw
Generowanie treści	Tworzenie nowych, koherentnych artefaktów informacyjnych	Produkowanie dokumentacji, kodu, materiałów marketingowych

Źródło: Opracowanie własne

Definicja funkcjonalna ma istotną zaletę praktyczną: uniezależnia ocenę „inteligencji” systemu od ontologicznych debat o naturze świadomości, koncentrując się na mierzalnych efektach działania (Russell i Norvig, 2023).

1.2.3. Test Turinga jako kryterium odniesienia

Test Turinga, zaproponowany przez Alana Turinga w 1950 roku w artykule *Computing Machinery and Intelligence*, stanowi historycznie pierwszą systematyczną próbę operacjonalizacji pojęcia maszynowej inteligencji (Turing, 1950). Turing zaproponował „grę imitacyjną” (ang. *imitation game*), w której sędzia-człowiek prowadzi konwersację jednocześnie z człowiekiem i maszyną, nie wiedząc, który jest który. Jeśli sędzia nie jest w stanie niezawodnie rozróżnić maszyny od człowieka, maszyna zdaje test.

Test Turinga ma charakter **behawioralny i pragmatyczny**: nie pyta, czy maszyna „myśli” w ontologicznym sensie, lecz czy jej zachowanie jest **nieodróżnialne od zachowania inteligentnego** (Turing, 1950). Badania wskazują, że w wielu scenariuszach współczesne modele językowe są trudne do odróżnienia od człowieka, co rodzi pytania o adekwatność tradycyjnych kryteriów (Jones i Bergen, 2025).

Jednocześnie test Turinga podlega krytyce za:

- **czysto lingwistyczną definicję** inteligencji, pomijającą inne modalności (percepcję, motorykę, emocje);
- **brak wymogu rozumienia** — maszyna może symulować inteligentne zachowanie bez posiadania wewnętrznych reprezentacji znaczenia;
- **podatność na manipulację** — skupienie na oszukaniu sędziego zamiast na rzeczywistych kompetencjach (Searle, 1980).

1.2.4. Współczesne rozumienie AI

Ewolucja technologiczna przesunęła rozumienie AI od **systemów regułowych** o wąskim zastosowaniu (starsza AI: *expert systems*, drzewa decyzyjne) do **systemów uczących się**, generatywnych i adaptacyjnych. Współczesna analogia ilustruje tę transformację (tabela 1.2).

TABELA 1.2. Porównanie starszej AI z nowszą

WYMIAR	STARSZA AI (PRZED 2012)	NOWOCZESNA AI (PO 2017)
Architektura	Systemy regułowe, pojedyncze algorytmy	Głębokie sieci neuronowe, architektura Transformer
Zdolność uczenia	Brak — wymagała ręcznego programowania	Uczenie się z danych, adaptacja do kontekstu
Zakres	Jedna funkcja, wąskie zastosowanie	Uniwersalność, <i>transfer learning</i>
Generatywność	Ograniczona do predefiniowanych wzorców	Tworzenie nowych, koherentnych treści
Interakcja	Skomplikowane interfejsy, wymagana ekspertyza	Język naturalny jako interfejs uniwersalny

Źródło: Opracowanie własne

Ta transformacja ma fundamentalne znaczenie dla biznesu, bo sztuczna inteligencja przestała być domeną inżynierów i analityków danych, a stała się narzędziem dostępnym dla specjalistów z każdej dziedziny, komunikujących się z systemami w języku naturalnym (Chalmers i in., 2026).

1.3. Historia i przełomowe momenty w rozwoju AI

Rzeczywisty rozwój sztucznej inteligencji nie był procesem liniowym, lecz charakteryzował się cyklicznymi falami entuzjazmu i zastoju — tzw. „zimami AI” (ang. *AI winters*) oraz okresami przełomowych odkryć. Zrozumienie tej historii pozwala na lepsze osadzenie współczesnych narzędzi w kontekście długoterminowych trendów technologicznych i uniknięcie zarówno nadmiernego optymizmu, jak i pesymizmu co do możliwości AI.

1.3.1. Przed narodzinami AI: fundamenty teoretyczne (1940 – 1955)

Przed sformalizowaniem pola badań nad sztuczną inteligencją powstały kluczowe koncepcje teoretyczne i techniczne:

- **1943:** McCulloch i Pitts publikują model sztucznego neuronu — oparty na biologicznej inspiracji, stanowiący fundament późniejszych sieci neuronowych (McCulloch i Pitts, 1943).
- **1950:** Alan Turing publikuje artykuł *Computing Machinery and Intelligence*, w którym wprowadza test Turinga i zadaje fundamentalne pytanie: „Czy maszyny mogą myśleć?” (Turing, 1950).
- **1951:** Powstają pierwsze maszyny uczące się — Marvin Minsky i Dean Edmonds budują SNARC (*Stochastic Neural Analog Reinforcement Calculator*), pierwszą sieć neuronową zaimplementowaną sprzętowo.

1.3.2. Konferencja w Dartmouth i narodziny dyscypliny (1956)

Lato 1956 roku uznawane jest za symboliczny moment narodzin sztucznej inteligencji jako odrębnej dyscypliny naukowej. Konferencja w Dartmouth College, zorganizowana przez Johna McCarthy’ego, przyciągnęła grupę badaczy, którzy zdefiniowali agendę badań na kolejne dekady (McCarthy i in., 2006).

Uczestnicy konferencji — w tym McCarthy, Minsky, Rochester i Shannon — założyli później wiodące ośrodki badawcze AI na uniwersytetach MIT, Stanford i Carnegie Mellon. Ich wizja „inteligencji maszynowej” była jednak znacznie bardziej optymistyczna niż rzeczywistość technologiczna tamtego okresu. Przewidywano, że w ciągu dekady maszyny będą potrafiły tłumaczyć języki, rozwiązywać problemy matematyczne i grać w szachy na poziomie mistrzowskim — cele te zostały jednak osiągnięte z wieloletnim opóźnieniem (Crevier, 1993).

1.3.3. Pierwsza „zima AI” i symboliczne odrodzenie (1974 – 1980)

Lata 70. przyniosły rozczarowanie wczesnymi obietnicami. Ograniczenia mocy obliczeniowej i złożoność rzeczywistych problemów doprowadziły do pierwszej „zimy AI” — redukcji finansowania badań i sceptycyzmu wobec potencjału pola. Jednocześnie nastąpił rozwój **systemów ekspertowych** — programów opartych na regułach logicznych, symulujących wiedzę specjalistyczną w wąskich domenach (np. MYCIN w medycynie, DENDRAL w chemii).

1.3.4. Druga „zima AI” i powrót sieci neuronowych (1987 – 1993)

Drugi okres stagnacji związany był z komercyjnym fiaskiem systemów ekspertowych, które okazały się kruche, trudne w utrzymaniu i niezdolne do generalizacji. Przełom nastąpił dzięki:

- **algorytmowi wstecznej propagacji** (ang. *backpropagation*, 1986) — efektywnej metodzie trenowania wielowarstwowych sieci neuronowych (Rumelhart i in., 1986);
- **rosnącej mocy obliczeniowej** — dającej możliwość praktycznej implementacji architektur teoretycznie znanych od lat.

1.3.5. Przełomowe momenty współczesnej ery AI (1997 – 2026)

Chronologia w tabeli 1.3 prezentuje kluczowe wydarzenia, które ukształtowały współczesny krajobraz sztucznej inteligencji w biznesie.

TABELA 1.3. Kluczowe momenty w rozwoju AI w latach 1997 – 2026

ROK	WYDARZENIE	ZNACZENIE
1997	IBM Deep Blue pokonuje Garry'ego Kasparowa w meczu szachowym (3,5:2,5)	Pierwszy globalny sygnał, że maszyny mogą przewyższyć człowieka w zadaniach strategicznych wymagających głębokiej analizy
2011	IBM Watson wygrywa w Jeopardy!	Demonstracja rozumienia języka naturalnego i przetwarzania nieustrukturyzowanej wiedzy (Ferrucci i in., 2010)
2012	AlexNet wygrywa ImageNet	Rewolucja Deep Learning: głębokie sieci konwolucyjne przewyższają tradycyjne metody w rozpoznawaniu obrazów o rząd wielkości (Krizhevsky i in., 2012)
2016	AlphaGo pokonuje Lee Sedola (4:1) w Go	Przełom w grach o niepełnej informacji i intuicyjnym charakterze; demonstracja siły <i>reinforcement learning</i> (Silver i in., 2016)
2017	Architektura Transformer („Attention Is All You Need“)	Fundament wszystkich współczesnych modeli językowych; mechanizm uwagi zastępuje rekurencję (Vaswani i in., 2017)
2020	GPT-3 (175 miliardów parametrów)	Model generujący tekst na poziomie zbliżonym do człowieka; udowodnienie skalowalności możliwości językowych (Brown i in., 2020)
2020	AlphaFold2	Rozwiązanie problemu związania białek — przełom w biologii strukturalnej i medycynie (Jumper i in., 2021)
2022	Publiczny dostęp do ChatGPT (30 listopada)	Najszybciej rosnąca aplikacja w historii internetu; demokratyzacja dostępu do generatywnej AI
2023–2026	Era modeli multimodalnych i MoE	Gemini, GPT, Claude, DeepSeek i inne — modele rozumiejące tekst, obraz, kod i dźwięk; architektura Mixture of Experts optymalizująca efektywność

Źródło: Opracowanie własne

1.3.6. Analiza wybranych kamieni milowych

IBM Deep Blue vs Garry Kasparow (1997)

Mecz szachowy z 1997 roku miał charakter symboliczny, przekraczający znaczenie samej gry. Kasparow, uważany za najlepszego mistrza wszech czasów, przegrał z systemem wykorzystującym **brute-force search** i specjalizowane układy VLSI do ewaluacji pozycji. Istotne były jednak też kontekstowe czynniki: presja czasowa, zmęczenie ludzkiego gracza kontra niezmiennosc maszyny oraz rola inżynierów IBM w dostosowywaniu systemu między partiami (Kasparow, 2021).

Kasparow w późniejszych analizach zwracał uwagę, że przegrana była „bliska” (3,5:2,5) i że Deep Blue wykazywało cechy sugerujące interwencję ludzką — co doprowadziło do debat o naturze „maszynowej” inteligencji.

Architektura Transformer (2017)

Opublikowana przez zespół Google Brain praca *Attention Is All You Need* (Vaswani i in., 2017) wprowadziła architekturę opartą wyłącznie na mechanizmie **uwagi** (ang. *attention*), eliminującą rekurencyjne połączenia charakterystyczne dla wcześniejszych sieci (LSTM, GRU).

Kluczowe innowacje w architekturze Transformer:

- **self-attention**: modelowanie zależności między wszystkimi elementami sekwencji równoległe;
- **multi-head attention**: równoległe projekcje uwagi pozwalające na wychwycenie różnych typów relacji;
- **position encoding**: kodowanie informacji o pozycji bez rekurencji.

Architektura ta stała się fundamentem **GPT** (*Generative Pre-trained Transformer*) i **BERT**, a przez to wszystkich współczesnych modeli językowych.

ChatGPT (2022)

Uruchomienie ChatGPT 30 listopada 2022 roku przez OpenAI oznaczało przejście z **playgroundów** i **API** dostępnych dla programistów do **interfejsu konwersacyjnego** zrozumiałego dla każdego użytkownika. W ciągu dwóch miesięcy aplikacja przekroczyła 100 milionów użytkowników i stała się najszybciej rosnącym produktem cyfrowym w historii (K. Hu, 2023).

Ten moment był przełomowy nie technologicznie (GPT-3 był znany od 2020 roku), lecz **interfejsowo i społecznie**, bo pokazał, że granica między „narzędziem dla ekspertów” a „infrastrukturą powszechnego użytku” może zostać przekroczona w ciągu jednej nocy.

1.3.7. Perspektywa historyczna

Historia AI wykazuje **cykliczność**: okresy nadmiernych oczekiwań, momenty rozczarowań, a następnie odrodzenie na wyższym poziomie technologicznym. Jednak tempo zmian po 2012 roku sugeruje **akcelerację** — czas między fundamentalnym odkryciem a zastosowaniem komercyjnym skraca się z dekad do miesięcy.

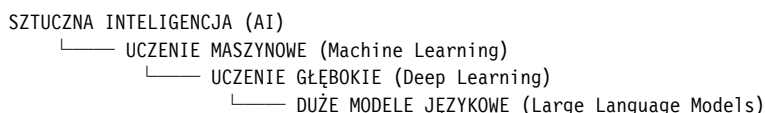
Warto zachować krytyczną perspektywę: wiele współczesnych „przełomów” to w rzeczywistości **skalowanie i inżynieria** znanych koncepcji, nie zaś fundamentalnie nowe paradygmaty. Architektura Transformer ma już blisko dekadę; kluczową zmianą jest skala danych, mocy obliczeniowej i zasobów inwestycyjnych.

1.4. Fundament technologiczny sztucznej inteligencji

Zrozumienie hierarchii pojęć oraz relacji między sztuczną inteligencją, uczeniem maszynowym, głębokim uczeniem i dużymi modelami językowymi jest kluczowe dla świadomego wyboru narzędzi biznesowych. Często spotyka się używanie tych terminów zamiennie, co prowadzi do błędnych oczekiwań co do możliwości i ograniczeń konkretnych rozwiązań.

1.4.1. Hierarchia pojęć: $AI \supset ML \supset DL \supset LLM$

Relację między czterema kluczowymi pojęciami można przedstawić jako zagnieżdżone zbiory, gdzie każdy poziom stanowi specjalizację poprzedniego:



Sztuczna inteligencja (AI) to najszersza kategoria, obejmująca wszystkie systemy wykazujące zachowania przypominające inteligencję — od prostych systemów regułowych po zaawansowane sieci neuronowe. AI nie wymaga uczenia się z danych; klasyczne systemy ekspertowe były programowane ręcznie przez specjalistów (Russell i Norvig, 2023).

Uczenie maszynowe (ML) to podzbiór AI, w którym systemy uczą się wzorców z danych, a nie są programowane. ML obejmuje zarówno proste algorytmy (regresja liniowa, drzewa decyzyjne), jak i zaawansowane metody.

Uczenie głębokie (DL) to technika ML wykorzystująca wielowarstwowe sieci neuronowe (stąd „głębokie”) do automatycznej ekstrakcji cech z surowych danych. Eliminuje to potrzebę ręcznej inżynierii cech, która była wąskim gardłem tradycyjnego ML (LeCun i in., 2015).

Duże modele językowe (LLM) to specyficzna klasa modeli głębokiego uczenia, zaprojektowana do przetwarzania i generowania języka naturalnego, charakteryzująca się dużą skalą parametrów (miliardy) i architekturą Transformer (Brown i in., 2020).

Analogia ilustrująca tę hierarchię:

- **AI** = wszystkie pojazdy (samochody, ciężarówki, motocykle, statki)
- **ML** = pojazdy osobowe
- **DL** = pojazdy osobowe elektryczne
- **LLM** = konkretny model, np. Tesla Model S

1.4.2. Uczenie maszynowe: trzy paradygmaty

Współczesne uczenie maszynowe opiera się na trzech głównych paradygmatach, różniących się dostępnością etykiet danych i sposobem uczenia (tabela 1.4).

TABELA 1.4. Paradygmaty uczenia maszynowego

PARADYGMAT	DANE	MECHANIZM UCZENIA	ANALOGIA BIZNESOWA
Uczenie nadzorowane (ang. <i>supervised learning</i>)	Dane z etykietami (znane poprawne odpowiedzi)	Model uczy się mapowania wejść na wyjścia na podstawie przykładów	Nowy pracownik z trenerem, który pokazuje poprawne odpowiedzi
Uczenie nienadzorowane (ang. <i>unsupervised learning</i>)	Dane bez etykiet	Model samodzielnie odkrywa struktury i wzorce w danych	Pracownik samodzielnie analizujący dokumentację firmy w poszukiwaniu regularności
Uczenie ze wzmocnieniem (ang. <i>reinforcement learning</i>)	Brak danych <i>a priori</i> ; informacja zwrotna ze środowiska	Agent uczy się poprzez próbę i błąd, za które otrzymuje odpowiednio nagrody lub kary	Pracownik podejmujący decyzje i uczący się na podstawie premii lub braku nagrody

Źródło: Opracowanie własne

Uczenie nadzorowane

W uczeniu nadzorowanym model otrzymuje pary (wejście, wyjście) i uczy się funkcji przekształcającej wejścia w wyjścia. Przykład zastosowania biznesowego: **wykrywanie oszustw finansowych** — model trenowany na historycznych transakcjach oznaczonych jako „prawidłowe” lub „oszustwo” uczy się rozpoznawać wzorce w nowych transakcjach.

Kluczowe wyzwanie polega na tym, że jakość modelu zależy od jakości i reprezentatywności danych treningowych. Bias w danych przekłada się bezpośrednio na bias w decyzjach modelu.

Uczenie nienadzorowane

Uczenie nienadzorowane stosowane jest wtedy, gdy etykiety są niedostępne lub zbyt kosztowne do zebrania. Model eksploruje dane w poszukiwaniu ukrytej struktury. Przykład: **segmentacja klientów** w e-commerce — model samodzielnie grupuje klientów według częstotliwości zakupów, wartości koszyka i zachowań przeglądarkowych bez wcześniejszego zdefiniowania kategorii.

Techniki obejmują: klasteryzację (K-means, DBSCAN), redukcję wymiarowości (PCA, t-SNE), oraz modele generatywne (GAN, VAE).

Uczenie ze wzmocnieniem

Agent uczy się sekwencyjnego podejmowania decyzji w środowisku, maksymalizując skumulowaną nagrodę. Nie otrzymuje poprawnych odpowiedzi, lecz informację zwrotną o jakości swoich działań.

Przykład biznesowy: **optymalizacja cen dynamicznych** — system testuje różne poziomy cen, obserwuje reakcję rynku (konwersję, przychód) i uczy się strategii cenowej maksymalizującej zysk.

W kontekście LLM, uczenie ze wzmocnieniem odgrywa kluczową rolę w **RLHF** (*Reinforcement Learning from Human Feedback*) — etapie dostrajania modeli do preferencji ludzkich (sekcja 1.6.4).

1.4.3. Deep Learning: wielowarstwowe reprezentacje

Uczenie głębokie opiera się na **sztucznych sieciach neuronowych** — modelach obliczeniowych inspirowanych biologicznymi neuronami, choć niebędących ich wiernym odwzorowaniem. Kluczowe cechy odróżniające DL od płytkich metod ML:

- **Hierarchiczna ekstrakcja cech:** W tradycyjnym ML analityk musiał ręcznie zdefiniować cechy (np. dla rozpoznawania twarzy: odległość między oczami, szerokość nosa). W DL pierwsze warstwy sieci uczą się wykrywać proste wzorce (krawędzie, tekstury), kolejne — ich kombinacje (kształty), a najgłębsze — semantyczne koncepty (twarze, obiekty) (LeCun i in., 2015).
- **End-to-end learning:** Sieć uczy się bezpośredniego mapowania z surowych danych (pikseli, fal dźwiękowych, znaków) na dane wyjściowe, eliminując potrzebę tworzenia pośrednich reprezentacji zaprojektowanych przez człowieka.
- **Skalowalność:** Wydajność modeli DL rośnie wraz z dostępnymi danymi i mocą obliczeniową — w przeciwieństwie do tradycyjnych algorytmów, które po osiągnięciu pewnej skali przestają się poprawiać.

Architektury głębokiego uczenia obejmują:

- **sieci konwolucyjne (CNN):** wykorzystywane głównie do przetwarzania obrazów;
- **sieci rekurencyjne (RNN, LSTM, GRU):** historycznie stosowane w sekwencjach (tekst, dźwięk), a obecnie wypierane przez Transformer;
- **sieci typu Transformer:** obecnie dominująca architektura dla języka i multimodalności.

1.4.4. Architektura Transformer

Opublikowana w 2017 roku przez zespół Google Brain praca *Attention Is All You Need* (Vaswani i in., 2017) wprowadziła architekturę, która zrewolucjonizowała przetwarzanie języka naturalnego i stała się fundamentem wszystkich współczesnych LLM.

Struktura oryginalnego Transformera

Oryginalny Transformer składał się z dwóch modułów:

- **encoder**: przekształca sekwencję wejściową w wektorową reprezentację kontekstową;
- **decoder**: generuje sekwencję wyjściową na podstawie reprezentacji enkodera.

Zadaniem było **maszynowe tłumaczenie**: encoder przetwarzał zdanie w języku źródłowym, decoder generował tłumaczenie w języku docelowym.

Mechanizm uwagi

Kluczową innowacją był mechanizm **self-attention**, pozwalający modelowi ważyć znaczenie różnych elementów sekwencji względem siebie:

„Dla każdego słowa w zdaniu model oblicza, na ile powinien »zwracać uwagę« na inne słowa, aby poprawnie je zinterpretować”. (Vaswani i in., 2017)

Przykład: w zdaniu „On położył książkę na stole, ponieważ była ciężka” — słowo „była” musi nawiązywać do „książkę”, nie do „stole”. *Self-attention* pozwala modelowi bezpośrednio modelować te zależności, niezależnie od odległości w sekwencji.

Zalety względem poprzednich architektur:

- **równoległość**: wszystkie pozycje przetwarzane są jednocześnie, a nie sekwencyjnie jak w RNN;
- **długodystansowy zasięg**: zależności między dowolnymi elementami sekwencji są modelowane bezpośrednio;
- **interpretowalność**: wagi uwagi pokazują, które fragmenty wejścia wpływają na konkretne wyjście.

Warianty architektoniczne

Obecnie modele językowe wykorzystują częściej tylko tryb dekodera. Tabela 1.5 przedstawia popularne modele i wykorzystane tryby.

TABELA 1.5. Zastosowanie wariantów architektury Transformer w wybranych modelach

WARIANT	STRUKTURA	REPREZENTATYWNE MODELE	ZASTOSOWANIE
Encoder-only	Tylko encoder	BERT, RoBERTa	Klasyfikacja tekstu, ekstrakcja informacji, zrozumienie języka
Decoder-only	Tylko decoder	GPT, LLaMA, Claude	Generowanie tekstu, autoregresyjne modelowanie
Encoder-decoder	Oba moduły	T5, BART, FLAN-T5	Tłumaczenie maszynowe, podsumowanie, parafraza

Źródło: Opracowanie własne

Współczesne LLM (GPT, Claude, Gemini) wykorzystują wariant **decoder-only**, który okazał się najskuteczniejszy dla generatywnych zadań językowych przy skalowaniu.

Wizualizacja skali

Strona <https://bbycroft.net/llm> oferuje interaktywną wizualizację architektury Transformer na różnych skalach (tabela 1.6).

TABELA 1.6. Wizualizowane modele GPT

MODEL	PARAMETRY	CHARAKTERYSTYKA
nano-GPT	85 584	Minimalna działająca implementacja edukacyjna
GPT-2 small	124 miliony	Pierwszy publicznie znaczący model generatywny
GPT-2 XL	1,5 miliarda	Demonstracja skalowalności możliwości
GPT-3	175 miliardów	Próg „emergentnych” zdolności <i>few-shot</i>

Źródło: Opracowanie własne na podstawie <https://bbycroft.net/llm>

Ta wizualizacja ilustruje, że architektura pozostaje fundamentalnie taka sama — zmienia się wyłącznie skala (liczba warstw, wymiarów, głowic uwagi), co prowadzi do jakościowo nowych zdolności.

Podobne wizualne wyjaśnienie, jak działa architektura Transformer, z możliwością dostosowania parametrów, jest udostępnione pod adresem <https://poloclub.github.io/transformer-explainer> (Cho i in., 2025).

1.4.5. Implikacje dla biznesu

Zrozumienie fundamentów technologicznych sztucznej inteligencji ma praktyczne konsekwencje dla decyzji biznesowych.

- Wybór narzędzia:** Prosty problem klasyfikacyjny może wymagać tylko klasycznego ML; generowanie tekstu wymaga LLM; zadanie specyficzne może wymagać fine-tuningu.
- Ocena kosztów:** Modele większe (mające więcej parametrów) nie zawsze są lepsze dla konkretnego zadania; często wystarczy mniejszy, dostrojony model.
- Zarządzanie oczekiwaniami:** LLM nie „rozumieją” w ludzkim sensie — przewidują statystycznie prawdopodobne sekwencje (sekcja 1.7.1).
- Strategia danych:** Jakość danych treningowych jest często ważniejsza niż architektura modelu — *garbage in, garbage out* obowiązuje na wszystkich poziomach hierarchii.

1.5. Jak działają duże modele językowe (LLM)?

Duże modele językowe stanowią obecnie najbardziej widoczny i szeroko adoptowany wariant sztucznej inteligencji w kontekście biznesowym. Zrozumienie mechanizmów ich działania — od poziomu tokenów po architekturę wnioskowania — jest kluczowe dla efektywnego wykorzystania tych narzędzi oraz świadomego zarządzania ich ograniczeniami.

1.5.1. Przewidywanie kolejnego tokenu

Fundamentalna zasada działania LLM jest zaskakująco prosta: **model przewiduje kolejny token w sekwencji**. Nie „rozumie” tekstu w ontologicznym sensie, lecz oblicza statystycznie najbardziej prawdopodobną kontynuację.

Proces ten można zilustrować przykładami (tabela 1.7).

TABELA 1.7. Przykłady przewidywania kolejnego tokenu

KONTEKST	PRAWDOPODOBNA KONTYNUACJA	MNIEJ PRAWDOPODOBNA
„Warszawa jest stolicą...”	„Polski”	„mazowieckiego”, „Europy”
„Woda wrze w temperaturze...”	„100 stopni Celsjusza”	„20 stopni”, „to zależy od ciśnienia”
„Po deszczu często pojawia się...”	„tęcza”	„susza”, „parasol”

Źródło: Opracowanie własne

W przypadku zdań o wysokiej przewidywalności (fakty encyklopedyczne, idiomy) model generuje niemal deterministyczne odpowiedzi. W kontekstach otwartych, kreatywnych lub niejednoznacznych — rozkład prawdopodobieństwa jest szerszy, co pozwala na generowanie bardziej zróżnicowanych kontynuacji.

Krytyczna konsekwencja jest taka, że **model nie posiada wiedzy w ludzkim sensie**. „Wie” jedynie, jakie sekwencje tokenów występują często w zbiorze treningowym. To rozróżnienie ma fundamentalne znaczenie dla oceny wiarygodności odpowiedzi LLM (sekcja 1.7.1).

1.5.2. Token: podstawowa jednostka przetwarzania

Token to podstawowa jednostka tekstu przetwarzana przez model. W przeciwieństwie do intuicyjnego rozumienia pojęcia „słowo”, tokenizacja jest zoptymalizowana pod kątem efektywności działania modelu, a nie czytelności dla człowieka.

Charakterystyka tokenów:

- Mogą reprezentować **całe słowa** (np. „the”, „and” w języku angielskim).
- Mogą także reprezentować **fragmenty słów** (podśłowa, morfemy): „przewidywanie” → [„przewid”, „ywanie”].
- **Znaki interpunkcyjne** to osobne tokeny.
- **Pojedyncze znaki** mogą być tokenami dla rzadkich słów lub symboli.

Tabela 1.8 przedstawia przybliżone przykłady tokenizacji.

Różne modele używają różnych tokenizatorów. GPT-3/4 używają BPE (*Byte Pair Encoding*), inne modele — SentencePiece lub podobnych algorytmów. To ma praktyczne konsekwencje, bo ten sam tekst może zajmować różną liczbę tokenów w różnych modelach, co wpływa na koszt i wykorzystanie okna kontekstowego.

TABELA 1.8. Przykłady liczby tokenów

TEKST	TOKENY (PRZYBLIŻONA LICZBA)
„AI zmienia biznes”	~4-5 tokenów
„Sztuczna inteligencja transformuje gospodarkę”	~8-10 tokenów
„Antyestablishmentarianizm”	~5-6 tokenów (jako rzadkie słowo)

Źródło: Opracowanie własne

1.5.3. Prompt: interfejs komunikacji z modelem

Prompt to tekst wejściowy, zapytanie lub instrukcja podana modelowi w celu nakierowania go na pożądaną odpowiedź. W przeciwieństwie do tradycyjnego programowania, w którym określa się precyzyjne kroki algorytmu, w interakcji z LLM komunikuje się **cel** w języku naturalnym.

Efektywność promptu zależy od:

- **jasności instrukcji** — stosowania konkretnych, nieambiwalentnych sformułowań;
- **kontekstu** — dostarczenia modelowi niezbędnych informacji tła;
- **formatu** — określenia oczekiwanej struktury odpowiedzi (np. JSON, lista, esej);
- **przykładów** (ang. *few-shot prompting*) — przedstawienia wzorca pożądanej odpowiedzi.

Przykład różnicy w jakości promptu:

SŁABY PROMPT	LEPSZY PROMPT
„Napisz o marketingu”	„Przygotuj 5-punktowy plan marketingowy dla start-upu SaaS B2B, skupiając się na niskobudżetowych kanałach cyfrowych. Użyj formatu: [Kanał] — [Działanie] — [Oczekiwany wynik].”

Prompt systemowy

Prompt systemowy (*system prompt*, *system message*) to zestaw instrukcji konfiguracyjnych przekazywanych modelowi **przed** interakcją z użytkownikiem. Definiuje on rolę, ograniczenia i styl odpowiedzi modelu.

Przykład promptu systemowego ChatGPT-5:

„Jesteś ChatGPT, dużym modelem językowym wytrenowanym przez OpenAI. Twoja baza wiedzy obejmuje dane do czerwca 2024...” (<https://github.com/elder-plinius/CL4R1T4S/blob/main/OPENAI/ChatGPT5-08-07-2025.mkd>).

Prompt systemowy jest niewidoczny dla użytkownika końcowego, ale filtruje i kształtuje wszystkie interakcje. Umożliwia dostosowanie modelu do konkretnych zastosowań biznesowych bez konieczności re-treningu.

1.5.4. Kontekst i okno kontekstowe

Kontekst to wszystkie informacje, które model „pamięta” w danej interakcji — analogicznie do pamięci krótkotrwałej lub roboczej. Obejmuje:

- aktualny prompt użytkownika;
- historię rozmowy (w przypadku interfejsów konwersacyjnych);
- prompt systemowy;
- przesłane pliki lub dokumenty.

Okno kontekstowe (ang. *context window*) to maksymalna liczba tokenów, którą model może przetworzyć w jednej interakcji. Po przekroczeniu tego limitu model „zapomina” najstarsze fragmenty — zazwyczaj przez mechanizm *sliding window* lub podobny.

Ewolucja okien kontekstowych w wybranych modelach (tabela 1.9).

TABELA 1.9. Przykłady rozmiaru okien kontekstowych

MODEL	OKNO KONTEKSTOWE	EKWIWALENT
GPT-3 (2020)	2048 tokenów	~1500 słów
GPT-4 (2023)	8192 / 32 768 tokenów	~6000 / 24 000 słów
Claude 3 (2024)	200 000 tokenów	~150 000 słów
Gemini 1.5 Pro (2024)	1 000 000 tokenów	~750 000 słów, ~30 000 linii kodu
GPT-4 Turbo (2024)	128 000 tokenów	~96 000 słów

Źródło: Opracowanie własne

Wzrost rozmiaru okien kontekstowych ma fundamentalne znaczenie biznesowe, bo umożliwia analizę całych raportów, kontraktów, baz kodu czy książek „za jednym razem”, bez konieczności fragmentacji i agregacji.

1.5.5. Parametry i wagi: wewnętrzna reprezentacja wiedzy

Parametry to wewnętrzne, nauczone wartości modelu — liczby zmiennoprzecinkowe, które definiują, jak model przetwarza wejście w wyjście. Współczesne LLM liczą parametry w **miliardach** (Brown i in., 2020). Tabela 1.10 prezentuje liczby parametrów w popularnych modelach.

TABELA 1.10. Modele językowe i deklarowana liczba parametrów

MODEL	PARAMETRY	UWAGI
GPT-1	117 milionów	Pierwszy model GPT, <i>proof-of-concept</i>
GPT-2	1,5 miliarda	Wstrzymana publikacja z obawy przed nadużyciami
GPT-3	175 miliardów	Próg „emergentnych” zdolności <i>few-shot</i>
GPT-4	~1 bilion (szacunek)	Multimodalność, nie ujawniono dokładnej liczby
LLaMA 3.1	405 miliardów	Największy otwarty model Meta
DeepSeek-V3	671 miliardów (MoE)	Aktywowane ~37 miliardów na token

Źródło: Opracowanie własne

Wagi (*weights*) to konkretne parametry określające siłę połączeń między neuronami. W architekturze Transformer wagi realizowane są jako macierze w operacjach liniowych (mnożenia macierz-wektor). To w wagach „siedzi” to, czego model nauczył się z danych treningowych — wzorce statystyczne, składniowe, semantyczne i faktyczne.

Istotne rozróżnienie polega na tym, że **więcej parametrów ≠ więcej wiedzy**. Parametry kodują zarówno „wiedzę” (fakty, relacje), jak i „umiejętności” (rozumienie języka, formatowanie, styl). Duży model może być „głębszy” w konkretnej domenie niż mniejszy model wyspecjalizowany w danym obszarze.

1.5.6. Temperatura: kontrola kreatywności

Temperatura to parametr generowania kontrolujący „odwagę” losowania kolejnych tokenów. Matematycznie to skalowanie logitów (surowych wyników modelu) przed funkcją softmax. Tabela 1.11 prezentuje zakresy temperatury wraz z wpływem na dobór tokenu.

TABELA 1.11. Zakresy temperatury w modelach językowych

ZAKRES TEMPERATURY	EFEKT	ZASTOSOWANIE
0,0 – 0,3	Rozkład „ostry”, prawie deterministyczny. Model wybiera najbardziej prawdopodobny token.	Fakty, kody, definicje — gdzie precyzja jest kluczowa
0,7 – 1,0	Kompromis między spójnością a różnorodnością.	Ogólne zastosowania, konwersacja
1,2 – 2,0	Rozkład „płaski”, częstszy wybór mniej prawdopodobnych tokenów.	Kreatywne pisanie, burza mózgów, eksploracja

Źródło: Opracowanie własne

Przykład wpływu temperatury (ten sam prompt, różne wartości):

Prompt: „Jutro rano zjem...”

T = 0,1	T = 0,7	T = 1,5
„śniadanie”.	„owsiankę z owocami”.	„zielone jabłko, które spadło z drzewa sąsiadów”.

Wysoka temperatura zwiększa ryzyko **halucynacji** — generowania treści pozornie sensownych, lecz niezgodnych z faktami (sekcja 1.9.2).

1.5.7. Mixture of Experts

Mixture of Experts (MoE) to architektura, w której model dzielony jest na mniejsze, wyspecjalizowane podsieci („ekspertów”). Dla każdego tokenu system **routingu** wybiera podzbiór ekspertów do aktywacji, zamiast przetwarzać całą sieć (Shazeer i in., 2017).

Zalety architektury MoE:

- **efektywność obliczeniowa:** przy 8 ekspertach i aktywacji 2 na token, koszt wnioskowania jest zbliżony do modelu 2 – 3 razy mniejszego;
- **skalowalność:** możliwość zwiększania liczby parametrów bez proporcjonalnego wzrostu kosztów;
- **specjalizacja:** eksperci mogą uczyć się różnych aspektów języka (składnia, fakty, kod, różne języki naturalne).

Przykłady modeli MoE:

- **Mixtral 8x7B:** 8 ekspertów po 7 miliardów parametrów, ~13 miliardów aktywnych na token;
- **DeepSeek-V3:** 256 ekspertów, 671 miliardów parametrów łącznie, ~37 miliardów aktywnych;
- **GPT-4** (prawdopodobnie): szacunki sugerują architekturę MoE, choć OpenAI nie potwierdza szczegółów.

Wyzwaniem MoE jest *load balancing*, czyli zapewnienie równomiernego wykorzystania ekspertów, aby uniknąć „przepalania” niektórych i niedouczenia innych.

1.5.8. RAG: wzbogacanie modelu o zewnętrzną wiedzę

RAG (*Retrieval-Augmented Generation*) to architektura łącząca LLM z bazą wiedzy, pozwalająca na generowanie odpowiedzi opartych na faktach spoza danych treningowych modelu (Lewis i in., 2020).

Proces RAG składa się z następujących elementów:

1. **Retrieval:** zapytanie użytkownika przekształcane jest w embedding (wektor semantyczny), a następnie wyszukiwane są najbardziej podobne dokumenty z bazy wiedzy;
2. **Augmentation:** pobrane dokumenty dołączane są do promptu jako kontekst;
3. **Generation:** model generuje odpowiedź opartą zarówno na swojej wewnętrznej wiedzy, jak i na dostarczonych dokumentach.

Zastosowania biznesowe RAG to:

- **chatboty korporacyjne** odpowiadające na podstawie wewnętrznej dokumentacji;
- **asystenci prawni** wyszukujący precedensy i regulacje;
- **wsparcie techniczne** z aktualną bazą wiedzy produktowej.

Przykład wdrożenia to **Gemini Notebook** (wcześniej NotebookLM) firmy Google, który pozwala użytkownikom „rozmawiać” z własnymi dokumentami oraz generuje odpowiedzi z odwołaniami do konkretnych źródeł.

PROGRAM PARTNERSKI

— GRUPY HELION —



1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ

Zmień swoją stronę WWW w działający bankomat!

Dowiedz się więcej i dołącz już dzisiaj!

<http://program-partnerski.helion.pl>

GRUPA
Helion 

Czy wiesz, do czego możesz zatrudnić AI we własnej firmie?

Sztuczna inteligencja, odmieniana przez wszystkie przypadki i konteksty, stała się w ostatnich latach jednym z najmocniej dyskutowanych tematów. Powstają liczne artykuły poświęcone fenomenowi AI, opasłe tomiszczą dotyczące jej rozwoju i publikacje odnoszące się do przyszłości, która nas w związku z tym czeka.

Jednak ten poradnik, choć nie unika niezbędnej teorii, jest zupełnie inny.

Artur Strzelecki stawia na praktykę — łączy podstawy technologiczne z kwestiami biznesowymi, pokazuje zarówno mechanizmy działania współczesnych systemów AI, jak i sposoby ich zastosowania w firmach i organizacjach. Za sprawą jego książki poznasz nie tylko możliwości poszczególnych narzędzi, lecz także związane z nimi ograniczenia, koszty, ryzyka, wymagania infrastrukturalne, konsekwencje organizacyjne i regulacyjne.

Dowiesz się między innymi:

- Czym właściwie jest AI
- Jakie są zasady efektywnego promptowania
- W jaki sposób można używać sztucznej inteligencji do pracy z tekstem, grafiką, plikami audio i wideo
- Jak efektywnie korzystać z asystentów i agentów AI
- W czym jeszcze — na poziomie operacyjnym, taktycznym i strategicznym — może pomóc sztuczna inteligencja

Artur Strzelecki

Jest profesorem Uniwersytetu Ekonomicznego w Katowicach i pracownikiem Katedry Informatyki, specjalizującym się w marketingu internetowym, analityce internetowej, wyszukiwarkach internetowych i zastosowaniach sztucznej inteligencji w biznesie. Czterokrotnie znalazł się w światowym rankingu Top 2% Scientist według Stanford University. Jest autorem licznych publikacji naukowych w renomowanych czasopismach międzynarodowych, a także uczestnikiem projektów badawczych i konferencji na kilku kontynentach. W działalności badawczej i dydaktycznej koncentruje się na wpływie AI na edukację, organizacje i biznes; prowadzi między innymi kurs *Narzędzia AI w biznesie*. Od ponad dekady współtworzy Festiwal SEO — największą w Polsce konferencję poświęconą optymalizacji stron internetowych — jak również współpracuje z przedsiębiorstwami w zakresie strategii widoczności online, analityki internetowej i marketingu cyfrowego.

onepress Wydawnictwo Naukowe **Helion**



Księgarnia internetowa:
onepress.pl



HELION S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel.: 32 230 98 63
onepress@onepress.pl

książkiklasybusiness

ebook dostępny na:

ebookpoint

ISBN 978-83-289-3986-8



9 788328 939868

Cena: 89,00 zł