

PJN - Lab 8 RAG prototype

Mikołaj Stefański

Etap 1

Zadanie 1 - RAG na piechotę

```
In [1]: from main_rag import rag_query

In [ ]: test_queries = [
    "Czym jest praca zespołowa?",
    "PAN",
    "LP-78",
    "Jak pokonać kryzys zaufania?",
    "Czy inflacja 2022 była wyższa niż 2021?",
    "Co Sinek mówił o liderach?",
    "Co to odpowiedzialność moralna?",
    "Co znajduje się pod klasztorem w Lubiniu?",
    "Jakie ogrzewanie jest projektowane dla branży przemysłowej?",
    "Co znajduje się pod Klasztorem Benedyktynów?"
]

print(f"Uruchamiam test RAG.\n")

for q in test_queries:
    result = rag_query(q, top_k=3, verbose=False)

    print('='*80)
    print(f"Zapytanie: {result['query']}")
    print(f"Przyjęta strategia: {result['strategy']}. Pewność semantyczna: {result['semantic_confidence']}")
    print('='*80)
    print(f"Odpowiedź modelu:\n{result['answer']}")
    print('='*80)

    print("Użyte źródła:")
    for src in result['sources']:
        clean_text = src['fragment'].replace('\n', ' ')[:100]
        print(f"  -> [Źródło ID: {src['id']}] {clean_text}...")

    print("\n\n")
```

Uruchamiam test RAG.

=====

Zapytanie: Czym jest praca zespołowa?

Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Pewność semantyczna: 0.80

=====

Odpowiedź modelu:

BRAK INFORMACJI W DOKUMENTACH

=====

Użyte źródła:

-> [Źródło ID: 882eddab-d6b3-4d5b-94b2-5279a10b0cd8] Ale czy działanie przedsiębiorstwa zgodnie z prawem, akceptowanymi zasadami, to źródło skutecznej pr...

-> [Źródło ID: 642df096-496e-46c7-8077-4749f8f4359d] Jeśli złamiesz prawo i zostaniesz zatrzymany przez policję, pamiętaj o swoich podstawowych prawach j...

-> [Źródło ID: c1544cc0-e2d8-4c18-9072-94340b6c2a56] Ogrzewanie elektryczne projektowane dla branży przemysłowej, czyli grzejniki elektryczne, czy nagrza...

=====

Zapytanie: PAN

Przyjęta strategia: Niska pewność semantyczna - ES boost.. Pewność semantyczna: 0.42

=====

Odpowiedź modelu:

[Źródło 1] , [Źródło 3]

=====

Użyte źródła:

-> [Źródło ID: 04c2fecd-d130-49ca-9cbc-127b2d306276] Wśród laureatów znajdują się m.in. prof. dr hab. Barbara Bilińska, członek korespondent PAN; dr hab....

-> [Źródło ID: 9c001923-ef8d-4906-8378-562f3b262b9c] Agnieszka Gałek Laureat Wojewódzkiego Konkursu Matematycznego: Czytaj więcej ODDZIAŁ PRZEDSZKOLNY Z ...

-> [Źródło ID: 8af4a73f-16ad-4c36-82db-4df32344c993] Włodzimierza Kołosa (chemia) Laureatem zostaje dr Bartosz Zamroń, pracownik Instytutu Chemii Organicznej

=====

Zapytanie: LP-78

Przyjęta strategia: Niska pewność semantyczna - ES boost.. Pewność semantyczna: 0.34

=====

Odpowiedź modelu:

BRAK INFORMACJI W DOKUMENTACH [Źródło 1]

=====

Użyte źródła:

-> [Źródło ID: 51022800-7e71-47ce-9ab6-dda94b75990d] Niepodległości 6 78-449 Oficjalna strona gminy Borne Sulinowo Według danych z 30 czerwca 2010 r. gmina...

-> [Źródło ID: 7d19d973-d938-49dd-af35-32a780bed75c] Dokonanie zmiany zarządzenia Dyrektora Generalnego LP nr 55 z dnia 24 września 2007 r. w części dotyczy...

-> [Źródło ID: a7fdf18a-4487-4f27-8e66-ff49cccd8dc76] Środki funduszu leśnego, w tym także funduszu leśnego lasów znajdujących się poza zarządem PGL LP, g...

Zapytanie: Jak pokonać kryzys zaufania?

Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Pewność semantyczna: 0.71

Odpowiedź modelu:

BRAK INFORMACJI W DOKUMENTACH

Użyte źródła:

- > [Źródło ID: 6052f69e-5944-494a-9176-60a3f8136de7] Klasztor Benedyktynów w Lubiniu posiada podziemny ciąg korytarzy, w którym pochowani zostali władcy ...
- > [Źródło ID: c2768e9e-5e91-498e-8ef8-56905dda4f47] Każdy zakamarek, który jest zbyt mały, wąski czy w inny sposób niedostępny dla tradycyjnych kołćówek...
- > [Źródło ID: e03f5057-1908-4b1e-bc1b-abe9016adbc2] Chciałabym zostać stewardessą, podróżować po świecie, zwiedzać kraje, poznawać różnorodne kultury - ...

Zapytanie: Czy inflacja 2022 była wyższa niż 2021?

Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Pewność semantyczna: 0.76

Odpowiedź modelu:

BRAK INFORMACJI W DOKUMENTACH. [Źródło 3]

Użyte źródła:

- > [Źródło ID: 9398bd62-fbac-4383-b36a-89d8af6dd34f] 2021 auta z niemiec na zamówienie Używane samochody z zagranicy Kiedy marzymy o własnym samochodzie, ...
- > [Źródło ID: 0fa23099-c77e-41d5-bc6f-979185d5171c] - 5 kwietnia, 2021 Prawidz iwie mysliwski choć bez śladu mresa ;0 admin - 6 kwietnia, 2021 Marlena K....
- > [Źródło ID: e894cff2-2e77-4aeb-87e6-b1a8feff65ee] Jeśli w dniach 29.04.-10.05.2022 będziesz się ze mną kontaktować lub złożysz zamówienie, odpowiem na...

Zapytanie: Co Sinek mówił o liderach?

Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Pewność semantyczna: 0.72

Odpowiedź modelu:

BRAK INFORMACJI W DOKUMENTACH

Użyte źródła:

- > [Źródło ID: e07b12c5-930b-491c-9ed2-d9add0db7279] Grzegorz Schetyna, lider PO: Platforma zawsze była ugrupowaniem centrowym, z mocnym filarem konserwatywnym, ...
- > [Źródło ID: e3181657-ba04-44f1-9c87-72d718b22064] Miesiąc temu dostałem w prezencie od Jacka Santorskiego bardzo ciekawą książkę Kevin Murray'a pt.: ...
- > [Źródło ID: c695fcfa5-f079-493f-bfd1-e686c3e02796] Gdybym doszedł do pierwszej chałupy, należącej właściwie do Pyzika Jacka, wybawcy naszego, zastałem wszyscy...)

Zapytanie: Co to odpowiedzialność moralna?

Przyjęta strategia: Niska pewność semantyczna - ES boost.. Pewność semantyczna: 0.64

=====

Odpowiedź modelu:

BRAK INFORMACJI W DOKUMENTACH

=====

Użyte źródła:

- > [Źródło ID: 3fa4e4a7-a9df-44f5-ab46-c75312ac7be3] Odpowiedzialność: Samostan owienie rozwija autonomię, co wiąże się z przejmowaniem odpowiedzialności ...
- > [Źródło ID: 24508201-83a3-4731-9c25-ef36ae0678c8] Pozdrawiam, Tomasz Grzyb P ewnie dlatego nie rozumiesz odpowiedzialności za słowo Za oszczeroство możn...
- > [Źródło ID: 1f853b5d-6cee-495f-9e6f-b7def809c00e] Jego sygnatariusze to bard zo podobna grupa - światowy, naukowo-kulturowy olimp, liberalna inteligenc...

=====

Zapytanie: Co znajduje się pod klasztorem w Lubiniu?

Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Pewność semantyczna: 0.73

=====

Odpowiedź modelu:

Opactwo w Lubiniu ma zachowaną szczegółową dokumentację źródłową sięgającą począt ków swojego powstania. [Źródło 1]

=====

Użyte źródła:

- > [Źródło ID: dd8f1723-e5c1-46c9-bbba-04b6689f73c3] Opactwo w Lubiniu, co unik alne w porównaniu z innymi zgromadzeniami w Polsce, ma zachowaną szczegółowo...
- > [Źródło ID: cf752316-5bbf-41fe-90b6-556e310e8984] Ma w sobie dwie najważniejsze rzeczy, które cechują lidera i widać to w tej rozmowie dobrze: spójn...
- > [Źródło ID: cd81ee2f-0909-4db2-a0c6-0262145b312f] Koktajl odżywczy na odchudzanie zastępujący tradycjne niezdrowe śniadania i kolacje.osób lubi to.Zna...

=====

Zapytanie: Jakie ogrzewanie jest projektowane dla branży przemysłowej?

Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Pewność semantyczna: 0.84

=====

Odpowiedź modelu:

Ogrzewanie elektryczne projektowane dla branży przemysłowej, czyli grzejniki elektryczne, czy nagrzewnice elektryczne. [Źródło 1]

=====

Użyte źródła:

- > [Źródło ID: c1544cc0-e2d8-4c18-9072-94340b6c2a56] Ogrzewanie elektryczne projektowane dla branży przemysłowej, czyli grzejniki elektryczne, czy nagrza...
- > [Źródło ID: 882eddab-d6b3-4d5b-94b2-5279a10b0cd8] Ale czy działanie przedsię biorstwa zgodnie z prawem, akceptowanymi zasadami, to źródło skutecznej pr...
- > [Źródło ID: 2c1fa667-febb-4ca7-b587-0fc9672443b] Inwestycja zlokalizowana jest przy ulicy Przeworskiej 1, w Dzielnicy Praga Południe w Warszawie, na ...

=====

Zapytanie: Co znajduje się pod Klasztorem Benedyktyńów?

Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Pewność semantyczna: 0.72

=====

Odpowiedź modelu:
BRAK INFORMACJI W DOKUMENTACH

=====
Użyte źródła:

- > [Źródło ID: 09a0747c-2ea3-4bb6-9d5e-59f902c05264] - Jesteśmy dumne z nagrody w konkursie organizowanym przez ARiMR pn. "Kubek, dzbanek czy makatka... rę..."
- > [Źródło ID: 39c5e04d-db90-47b2-9f94-aa4ffa752ddf] Istnieje zestaw elementów, które możesz z łatwością łączyć ze sobą - od motywów, które sprawiają, że...
- > [Źródło ID: 98ebb177-a456-43fa-98fa-929586889472] Kupując nowy komputer koniecznie zwróćmy uwagę na generację takich podzespołów jak procesor, pamięć ...

Wyniki

Testy wykazały, że skuteczność systemu RAG jest bezpośrednią wypadkową jakości retrievalu oraz zdolności modelu Gemma do pracy z kontekstem. System poprawnie obsłużył kluczowe scenariusze: pełny sukces, obsługa akronimów oraz detekcja braku wiedzy.

Ocena jakości wyników

1. Sanity check

Pytania o "ogrzewanie przemysłowe" oraz "podziemia klasztoru" zakończyły się pełnym sukcesem.

- retrieval: bezbłędnie wyłowił dokumenty źródłowe o grzejnikach i historii opactwa,
- generacja: model Gemma idealnie sparafrasował treść, np.: "Ogrzewanie elektryczne projektowane dla branży przemysłowej, czyli grzejniki elektryczne...",
- wniosek: pipeline działa jak należy - retrieval dostarcza fakty, a LLM ubiera je w język naturalny.

2. Sukces hybrydy

Dla zapytania "PAN", system automatycznie dobrał strategię ES boost. Dzięki temu nie szukał semantycznego znaczenia słowa "Pan", lecz konkretnego akronimu w ES,

- wynik: model poprawnie wymienił nazwiska naukowców z kontekstu i zacytował źródło [Źródło 1],
- wniosek: mechanizm dynamicznych wag z poprzednich laboratoriów jest krytyczny dla poprawnego działania RAG przy nazwach własnych i skrótach.

3. Analiza błędu retrievalu

Zaobserwowano ciekawe zjawisko niedeterminizmu wektorów.

- Przy zapytaniu abstrakcyjnym "Jak pokonać kryzys zaufania?", system znalazł dokument o Klasztorze Benedyktyńów jako szum semantyczny.

- przy zapytaniu konkretnym "Co znajduje się pod Klasztorem Benedyktyńów?"*, system początkowo nie znalazł tego dokumentu. Wynikało to z faktu, że **Confidence Score** ~0.72 był na tyle wysoki, by zostać przy Qdrancie, ale na tyle niski, by zgubić dokument.
- wniosek: w systemach produkcyjnych, wykrycie nazwy własnej powinno wymuszać Keyword Search niezależnie od wyniku modelu wektorowego.

4. Odporność na halucynacje

Dla zapytań o Simona Sineka czy inflację, gdzie bazy nie zawierały odpowiedzi, model konsekwentnie zwracał: "BRAK INFORMACJI W DOKUMENTACH". Jest to pożąданie zachowanie w systemach biznesowych, minimalizujące ryzyko wprowadzania użytkownika w błąd. Restrykcyjny prompt skutecznie zablokował tak zwaną "wiedzę własną" modelu.

Implementacja RAG oparta na modelu Gemma i wyszukiwaniu hybrydowym spełnia swoje zadanie. Kluczowym czynnikiem limitującym jakość odpowiedzi okazała się zawartość bazy danych (**culturax**), która jest zbiorem losowym, a nie encykopedycznym. Zastosowanie restrykcyjnego promptu wyeliminowało halucynacje, a chunking (150 słów) zapewnił modelowi wystarczający kontekst do sformułowania poprawnej gramatycznie odpowiedzi.

Zadanie 2

```
In [8]: from main_rag import rag_query

final_test_queries = [
    "Czym jest praca zespołowa?",
    "PAN",
    "LP-78",
    "Jak pokonać kryzys zaufania?",
    "Czy inflacja 2022 była wyższa niż 2021?",
    "Co Sinek mówił o liderach?",
    "Dąbrówka badania nad językiem",
    "Ludzie działający razem",
    "Jakie dokumenty o klimacie po 2020?",
    "Co to odpowiedzialność moralna?"
]

print(f"Rozpoczynam test na 10 zapytaniach.\n")

for i, query in enumerate(final_test_queries, 1):
    print(f"{'='*80}")
    print(f"Zapytanie #{i}/10: '{query}'")

    result = rag_query(query, top_k=3, verbose=False)

    print(f"Przyjęta strategia: {result['strategy']}. Confidence score: {result['confidence']}")
    print(f"{'='*80}")
    print(f"Odpowiedź modelu: {result['answer']}")
    print(f"{'='*80}")

    print("Źródła:")
    for src in result['sources']:
```

```
    clean_frag = src['fragment'][:100].replace('\n', ' ')
    print(f" -> [ID: {src['id']}]. Score: {src['score']:.2f}) {clean_frag}.
print("\n")
```

Rozpoczynam test na 10 zapytaniach.

=====

Zapytanie #1/10: 'Czym jest praca zespołowa?'

Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Confidence score: 0.80

Odpowiedź modelu: BRAK INFORMACJI W DOKUMENTACH

Źródła:

-> [ID: 882eddab-d6b3-4d5b-94b2-5279a10b0cd8]. Score: 5.00) Ale czy działanie przedsiębiorstwa zgodnie z prawem, akceptowanymi zasadami, to źródło skutecznej pr.

-> [ID: 642df096-496e-46c7-8077-4749f8f4359d]. Score: 2.50) Jeśli złamiesz prawo i zostaniesz zatrzymany przez policję, pamiętaj o swoich podstawowych prawach j.

-> [ID: c1544cc0-e2d8-4c18-9072-94340b6c2a56]. Score: 1.67) Ogrzewanie elektryczne projektowane dla branży przemysłowej, czyli grzejniki elektryczne, czy nagrz e.

=====

Zapytanie #2/10: 'PAN'

Przyjęta strategia: Niska pewność semantyczna - ES boost.. Confidence score: 0.42

Odpowiedź modelu: Prof. dr hab. Barbara Bilińska, członek korespondent PAN; dr hab. Jarosław Mederski z Instytutu Matematycznego PAN; dr inż. Maciej Radzieński z Instytutu Maszyn Przepływowych PAN; dr Bartosz Zambroń z Instytutu Chemii Organicznej PAN; zespół naukowy z Instytut Medycyny Doświadczalnej i Klinicznej im. [Źródło 1]

Źródła:

-> [ID: 04c2fecd-d130-49ca-9cbc-127b2d306276]. Score: 10.00) Wśród laureatów znajdują się m.in. prof. dr hab. Barbara Bilińska, członek korespondent PAN; dr hab.

-> [ID: 9c001923-ef8d-4906-8378-562f3b262b9c]. Score: 5.00) Agnieszka Gałek Laureat Wojewódzkiego Konkursu Matematycznego: Czytaj więcej ODDZIAŁ PRZEDSZKOLNY Z .

-> [ID: 8af4a73f-16ad-4c36-82db-4df32344c993]. Score: 3.33) Włodzimierza Kołosa (chemia) Laureatem zostaje dr Bartosz Zambroń, pracownik Instytutu Chemii Organi.

=====

Zapytanie #3/10: 'LP-78'

Przyjęta strategia: Niska pewność semantyczna - ES boost.. Confidence score: 0.34

Odpowiedź modelu: BRAK INFORMACJI W DOKUMENTACH [Źródło 1]

Źródła:

-> [ID: 51022800-7e71-47ce-9ab6-dda94b75990d]. Score: 10.00) Niepodległości 6 7 8-449 Oficjalna strona gminy Borne Sulinowo Według danych z 30 czerwca 2010 r. gmi.

-> [ID: 7d19d973-d938-49dd-af35-32a780bed75c]. Score: 5.00) Dokonanie zmiany zarządzenia Dyrektora Generalnego LP nr 55 z dnia 24 września 2007 r. w części dotyc.

-> [ID: a7fdf18a-4487-4f27-8e66-ff49cccd8dc76]. Score: 3.33) Środki funduszu leśnego, w tym także funduszu leśnego lasów znajdujących się poza zarządem PGL LP,

g.

=====

Zapytanie #4/10: 'Jak pokonać kryzys zaufania?'

Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Confidence score: 0.71

Odpowiedź modelu: BRAK INFORMACJI W DOKUMENTACH

Źródła:

- > [ID: 6052f69e-5944-494a-9176-60a3f8136de7]. Score: 5.00) Klasztor Benedyktynów w Lubiniu posiada podziemny ciąg korytarzy, w którym pochowani zostali władcy.
 - .
 - > [ID: c2768e9e-5e91-498e-8ef8-56905dda4f47]. Score: 2.50) Każdy zakamarek, który jest zbyt mały, wąski czy w inny sposób niedostępny dla tradycyjnych końcówek.
 - > [ID: e03f5057-1908-4b1e-bc1b-abe9016adbc2]. Score: 1.67) Chciałabym zostać stewardessą, podróżować po świecie, zwiedzać kraje, poznawać różnorodne kultury
 - .

=====

Zapytanie #5/10: 'Czy inflacja 2022 była wyższa niż 2021?'

Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Confidence score: 0.76

Odpowiedź modelu: BRAK INFORMACJI W DOKUMENTACH [Źródło 3]

Źródła:

- > [ID: 9398bd62-fbac-4383-b36a-89d8af6dd34f]. Score: 5.00) 2021 auta z niemiec na zamówienie Używane samochody z zagranicyKiedy marzymy o własnym samochodzie, .
 - > [ID: 0fa23099-c77e-41d5-bc6f-979185d5171c]. Score: 2.50) – 5 kwietnia, 2021 Prawidliwie mysliwski choć bez śladu mresa ;0 admin – 6 kwietnia, 2021 Marlena K..
 - > [ID: e894cff2-2e77-4aeb-87e6-b1a8feff65ee]. Score: 1.67) Jeśli w dniach 29.04.-10.05.2022 będziesz się ze mną kontaktować lub złożysz zamówienie, odpowiem na.

=====

Zapytanie #6/10: 'Co Sinek mówił o liderach?'

Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Confidence score: 0.72

Odpowiedź modelu: BRAK INFORMACJI W DOKUMENTACH

Źródła:

- > [ID: e07b12c5-930b-491c-9ed2-d9add0db7279]. Score: 5.00) Grzegorz Schetyna, lider PO: Platforma zawsze była ugrupowaniem centrowym, z mocnym filarem konserwata.
 - > [ID: e3181657-ba04-44f1-9c87-72d718b22064]. Score: 2.50) Miesiąc temu dostałem w prezencie od Jacka Santorskiego bardzo ciekawą książkę Kevina Murray'a pt.:
 - .
 - > [ID: c695fca5-f079-493f-bfd1-e686c3e02796]. Score: 1.67) Gdy doszedłem do pierwszej chałupy, należącej właściwie do Pyzika Jacka, wybawcy naszego, zastałem wszyscy.

=====

Zapytanie #7/10: 'Dąbrówka badania nad językiem'
Przyjęta strategia: Niska pewność semantyczna - ES boost.. Confidence score: 0.61

Odpowiedź modelu: BRAK INFORMACJI W DOKUMENTACH.
[Źródło 3]

=====

Źródła:

- > [ID: 21ce6e28-030d-45b7-b70f-75e7ab961e68]. Score: 10.00) Ostatnią, choć nie najmniej istotną korzyścią wynikającą z dwujęzyczności jest opóźniony proces stara.
- > [ID: d83ea0dc-1ee3-41e3-8966-f9d8303a0281]. Score: 5.00) Zgłaszaając się do BADANIA TOMOGRAFII KOMPUTEROWEJ Badanie Tomografii Komputerowej Głównej Sposób przyg.
- > [ID: 5c39676b-746e-4735-8645-e72e50bfff9ba]. Score: 3.33) Bardzo nowoczesną metodą badania cytogenetycznego jest badanie FISH, w przebiegu którego łączy się t.

=====

Zapytanie #8/10: 'Ludzie działający razem'
Przyjęta strategia: Wysoka pewność semantyczna - Qdrant boost.. Confidence score: 0.70

Odpowiedź modelu: BRAK INFORMACJI W DOKUMENTACH.

=====

Źródła:

- > [ID: 7cf4e217-b04d-4ab1-b9ae-4e614142e14d]. Score: 5.00) Uważacie, że to duże ułatwienie dla rodziców, czy może raczej ograniczamy w ten sposób dziecko i neg.
- > [ID: a294dbdb-fa43-4724-99fb-19f34d90415e]. Score: 2.50) Tego dnia dzieciaki uczęszczające do oddziału przedszkolnego działającego w tutejszej szkole podstaw.
- > [ID: 98e64bde-c406-435f-8877-3501c9295d69]. Score: 1.67) Możesz liczyć się z średnim wydatkiem rzędu 98 zł za dzień, ale nasi użytkownicy znaleźli w tym m.

=====

Zapytanie #9/10: 'Jakie dokumenty o klimacie po 2020?'

Przyjęta strategia: Niska pewność semantyczna - ES boost.. Confidence score: 0.62

Odpowiedź modelu: BRAK INFORMACJI W DOKUMENTACH [Źródło 1]

=====

Źródła:

- > [ID: 09f390f1-5dc5-4c15-bf30-a70a8da02621]. Score: 10.00) W praktyce zaś, kontrola krzyżowa odbywa się po uprzednim poinformowaniu podatnika o konieczności pr.
- > [ID: 6c9c6ed2-c4f1-42bb-bb9a-9bd7ec7ba523]. Score: 5.00) Jeżeli chodzi o Wilię Błażennego, żadnych dokumentów kanonizacyjnych nie ma, a w dokumentach, do.
- > [ID: 70a53fa5-15b3-49f8-9e49-744f4234960e]. Score: 3.33) Po kilku latach intensywnej pracy studyjnej nad muzyką innych artystów, stworzyli własną - intensywne.

=====

Zapytanie #10/10: 'Co to odpowiedzialność moralna?'

Przyjęta strategia: Niska pewność semantyczna - ES boost.. Confidence score: 0.64

Odpowiedź modelu: BRAK INFORMACJI W DOKUMENTACH

Źródła:

- > [ID: 3fa4e4a7-a9df-44f5-ab46-c75312ac7be3]. Score: 10.00) Odpowiedzialność Samostanowienie rozwija autonomię, co wiąże się z przejmowaniem odpowiedzialności.
- > [ID: 24508201-83a3-4731-9c25-ef36ae0678c8]. Score: 5.00) Pozdrawiam, Tomasz Grzyb Pewnie dlatego nie rozumiesz odpowiedzialności za słowo Za oszczerstwo może.
- > [ID: 1f853b5d-6cee-495f-9e6f-b7def809c00e]. Score: 3.67) Jego sygnatariusze to bardzo podobna grupa – światowy, naukowo-kulturowy olimp, liberalna inteligen c.

Tabela wyników

Zapytanie	Typ	Strategia	Odpowiedź Modelu	Analiza źródeł
Czym jest praca zespołowa?	Semantyczne	Qdrant (0.80)	BRAK INFORMACJI	System znalazł teksty o "działaniu przedsiębiorstwa" i "prawach". Tematyka organizacji jest bliska semantycznie, ale w bazie brakowało definicji, więc model słusznie odmówił odpowiedzi.
PAN	Akronim	ES (0.42)	SUKCES	Idealne przełączenie na słowa kluczowe. Źródło nr 1 zawierało rozwinięcie skrótu i nazwiska badaczy związanych z PAN. Model poprawnie zacytował źródło.
LP-78	ID dok.	ES (0.34)	BRAK INFORMACJI	System poprawnie wykrył słowa kluczowe "LP" i "78", ale konkretnego dokumentu o ID "LP-78" nie było w bazie. Prawidłowa odmowa.
Jak pokonać kryzys zaufania?	Abstrakcyjne	Qdrant (0.71)	BRAK INFORMACJI	Wystąpił szum informacyjny. Model wektorowy powiązał "kryzys/zaufanie" z "klasztorem" i "stewardessą". Brak trafnych dokumentów w kontekście.
Czy inflacja 2022 wyższa...?	Factual	Qdrant (0.76)	BRAK INFORMACJI	Znaleziono daty 2021 i 2022, ale w kontekście aut i zamówień, a nie ekonomii.

Zapytanie	Typ	Strategia	Odpowiedź Modelu	Analiza źródeł
Co Sinek mówił o liderach?	Filtr: autor	Qdrant (0.72)	BRAK INFORMACJI	Model słusznie odmówił porównania, nie znajdując danych ekonomicznych.
Dąbrówka badania nad językiem	Metadane	ES (0.61)	BRAK INFORMACJI	Znaleziono słowo "lider", ale brak nazwiska Sinek. System nie pomylił osób -- brak halucynacji.
Ludzie działający razem	Parafraza	Qdrant (0.70)	BRAK INFORMACJI	Hybryda znalazła tekst o "badaniach" i "języku", ale brakło spójnego dokumentu o Dąbrówce.
Dokumenty o klimacie po 2020	Dwuważkowe	ES (0.62)	BRAK INFORMACJI	Wektory powiązały to z "rodzicami", "dzieciakami" i "grupą przedszkolną". Bardzo blisko tematycznie, ale brak definicji współpracy.
Co to odpowiedzialność moralna?	Semantyczne	ES (0.64)	BRAK INFORMACJI	ES znalazł słowo "dokumenty", ale zignorował kontekst klimatyczny. Wektoryzacja nie wychwyciła tematu zmian klimatu w tym zbiorze.
				Źródło nr 1 ("Samostanowienie rozwija autonomię...") było bardzo bliskie znaczeniowo, ale restrykcyjny prompt powstrzymał model przed naciąganiem faktów.

Etap 2

Zadanie 2

```
In [9]: from main_rag import rag_query

test_query = "Jakie ogrzewanie jest projektowane dla branży przemysłowej?"

scenarios = [
    {"k": 1, "opis": "1 fragment; skupienie - ok. 150 słów"},
    {"k": 3, "opis": "3 fragmenty; standard - ok. 450 słów"},
    {"k": 5, "opis": "5 fragmentów; szum - ok. 750+ słów"
]

print(f"Rozpoczynam eksperyment z chunkingiem.\n")
print(f"Pytanie: {test_query}\n")

for scen in scenarios:
    k = scen['k']
```

```
print(f"{'#'*60}")
print(f"Scenariusz: {scen['opis']} (top_k={k})")
print(f"{'='*60}")

result = rag_query(test_query, top_k=k, verbose=False)

print(f"Odpowiedź modelu: {result['answer']}")
print(f"{'='*60}")

print("Dostarczone fragmenty:")
for i, src in enumerate(result['sources'], 1):
    preview = src['fragment'][:80].replace('\n', ' ')
    print(f" [{i}] Score: {src['score']:.2f} | {preview}...")
print("\n\n")
```

Rozpoczynam eksperyment z chunkingiem.

Pytanie: 'Jakie ogrzewanie jest projektowane dla branży przemysłowej?'

```
#####
Scenariusz: 1 fragment; skupienie - ok. 150 słów (top_k=1)
=====
```

Odpowiedź modelu: Ogrzewanie elektryczne projektowane dla branży przemysłowej to grzejniki elektryczne, czy nagrzewnice elektryczne [Źródło 1].

```
=====
Dostarczone fragmenty:
```

[1] Score: 6.00 | Ogrzewanie elektryczne projektowane dla branży przemysłowej, czyli grzejniki ele...

```
#####
Scenariusz: 3 fragmenty; standard - ok. 450 słów (top_k=3)
=====
```

Odpowiedź modelu: Grzejniki elektryczne, czy nagrzewnice elektryczne [Źródło 1].

```
=====
Dostarczone fragmenty:
```

[1] Score: 6.00 | Ogrzewanie elektryczne projektowane dla branży przemysłowej, czyli grzejniki ele...

[2] Score: 2.50 | Ale czy działanie przedsiębiorstwa zgodnie z prawem, akceptow anonymi zasadami, to ...

[3] Score: 1.67 | Inwestycja zlokalizowana jest przy ulicy Przeworskiej 1, w Dz ielnicy Praga Połud...

```
#####
Scenariusz: 5 fragmentów; szum - ok. 750+ słów (top_k=5)
=====
```

Odpowiedź modelu: Ogrzewanie elektryczne projektowane dla branży przemysłowej, czyli grzejniki elektryczne, czy nagrzewnice elektryczne. [Źródło 1]

```
=====
Dostarczone fragmenty:
```

[1] Score: 6.00 | Ogrzewanie elektryczne projektowane dla branży przemysłowej, czyli grzejniki ele...

[2] Score: 2.50 | Ale czy działanie przedsiębiorstwa zgodnie z prawem, akceptow anonymi zasadami, to ...

[3] Score: 1.67 | Inwestycja zlokalizowana jest przy ulicy Przeworskiej 1, w Dz ielnicy Praga Połud...

[4] Score: 1.25 | Ale już po Świątach, bo zbliża się teraz czas dla rodziny i p rzyjaciół, a nie dl...

[5] Score: 1.00 | Gremlin jest owszem i złośliwy, ale jest takim Gargamelem, kt óry w swojej złośli...

Wyniki

Celem było zbadanie odporności modelu na nadmiar informacji. Wykorzystano pytanie o "ogrzewanie przemysłowe", dla którego w bazie istniała jedna poprawna odpowiedź.

Sprawdzaliśmy, czy dodanie nieistotnych dokumentów (top_k=3 i top_k=5) zmyli model.

Scenariusz	Liczba fragmentów	Odpowiedź Modelu	Czy zacytował źródło?	Ocena stabilności
Skupienie	1 (ok. 150 słów)	"Ogrzewanie elektryczne... to grzejniki elektryczne..."	TAK [Źródło 1]	Bezbłędna. Model otrzymał tylko relevantny tekst, więc odpowiedź była precyzyjna.
Standard	3 (ok. 450 słów)	"Grzejniki elektryczne, czy nagrzewnice elektryczne"	TAK [Źródło 1]	Bezbłędna. Model zignorował fragmenty o "prawie przedsiębiorstw" (Źródło 2) i "ulicy Przeworskiej" (Źródło 3).
Szum	5 (ok. 750+ słów)	"Ogrzewanie elektryczne... to grzejniki elektryczne..."	TAK [Źródło 1]	Bezbłędna. Mimo wprowadzenia silnego szumu, model bezbłędnie wyekstrahował fakt z pierwszego dokumentu.

Wnioski

1. Odporność na szum

Model Gemma wykazał się wysoką zdolnością kupienia uwagi. Mimo że w scenariuszu `top_k=5` aż 80% kontekstu stanowiły informacje śmieciowe, model nie halucynował i nie pomieszał faktów.

2. Stabilność cytowania

Niezależnie od długości kontekstu, model konsekwentnie identyfikował [Źródło 1] jako jedynego dostarczyciela wiedzy. To kluczowa cecha dla budowania dobrych systemów RAG.

3. Optymalny chunking

Dla modelu Gemma2B okno kontekstowe wypełnione 5 fragmentami po ~150 słów nie powoduje spadku jakości odpowiedzi. Sugeruje to, że w produkcji można bezpiecznie używać `top_k=5`, aby zwiększyć szansę na znalezienie odpowiedzi, nie ryzykując utraty precyzji.

Etap 3

```
In [ ]: from main_rag import rag_query

test_queries = [
    "Co Sinek mówił o liderach?", # Test na halucynacje, bo brak wiedzy w bazie
    "Jakie ogrzewanie jest projektowane dla branży przemysłowej?" # Test poprawny
]

print(f"Porównanie promptów: prostego (A) i rozbudowanego (B). \n")
```

```
for q in test_queries:
    print(f"{'='*80}")
    print(f"Pytanie: '{q}'")
    print(f"{'='*80}")

    print(f"Wersja A - prosty prompt:")
    res_a = rag_query(q, top_k=3, prompt_mode="A", verbose=False)
    print(f"Odpowiedź:\n{res_a['answer']}\n")

    print(f"- {'='*40}")
    print(f"Wersja B - rozbudowany prompt")
    res_b = rag_query(q, top_k=3, prompt_mode="B", verbose=False)
    print(f"Odpowiedź:\n{res_b['answer']}\n")

print("\n\n")
```

Porównanie promptów: prostego (A) i rozbudowanego (B).

Pytanie: 'Co Sinek mówił o liderach?'

Wersja A - prosty prompt:

Odpowiedź:

Według fragmentu [Źródło 1] Grzegorz Schetyna opisuje Platformę Obywatelską (PO) jako ugrupowanie centrowego, z mocnym filarem konserwatywnym i liberalnym.

[Źródło 2] wspomina o książce "Język liderów" Kevina Murray'a, którą dostała Jack Santorskiego. [Źródło 3] opisuje, jak Sinek (Pyzika Jack) spotyka się z dżentlmenami w chałupie i nie jest to oparty na jego słowach, ale raczej o opisanym przez autora zdarzeniu.

Wersja B - rozbudowany prompt

Odpowiedź:

Sinek nie podał informacji na temat tego, co powiedział o liderach. Nie znalazłem informacji w bazie. [1]

Pytanie: 'Jakie ogrzewanie jest projektowane dla branży przemysłowej?'

Wersja A - prosty prompt:

Odpowiedź:

Ogrzewanie elektryczne projektowane dla branży przemysłowej to **grzejniki elektryczne** czy **nagrzewnice elektryczne**.

Wersja B - rozbudowany prompt

Odpowiedź:

W branży przemysłowej projektowane są grzejniki elektryczne i nagrzewnice elektryczne. [1]

Wyniki

Celem eksperymentu było sprawdzenie, jak prompty wpływają na skłonność modelu do halucynacji oraz w jaki sposób znajduje źródła do cytatów. Porównano 2 wersje promptu: prostą wersję A z rozbrdudowanym i restrykcyjnym promptem B.

Scenariusz	Prompt A (Prosty)	Prompt B (Restrykcyjny)	Wnioski
Brak wiedzy ("Co Sinek mówił...?")	HALUCYNACJA Model próbował za wszelką cenę udzielić odpowiedzi. Powiązał nazwisko "Pyzika"	SUKCES. Model zgodnie z instrukcją przyznał: "Nie znalazłem informacji	Prompt B jest bezpieczny. Prompt A generuje szum i wprowadza w błąd.

Scenariusz	Prompt A (Prosty)	Prompt B (Restrykcyjny)	Wnioski
	z zapytaniem o Sineka i zaserwował informacje o Grzegorzu Schetynie i Kevinie Murray'u, co jest mylące dla użytkownika.	w bazie". Zadziałała instrukcja negatywna.	
Fakt w bazie ("Ogrzewanie...")	Poprawna, ale bez cytatu. Model podał prawidłową definicję, ale zignorował wymóg podania źródła w nawiasie (brak [1]).	Poprawna z cytatem. Model podał definicję i zakończył ją przypisem [1], co pozwala na weryfikację.	Prompt B wymusza weryfikowalność odpowiedzi, co jest kluczowe w RAG.

Wnioski

Model językowy nawet tak mały, jak Gemma2B bez silnych ograniczeń w prompcie ma tendencję do "zadowalania użytkownika", co prowadzi do zmyślania faktów, gdy brakuje wiedzy w kontekście. Zastosowanie restrykcyjnego promptu (wersja B) z jasnymi zasadami to znaczy: "jeśli nie wiesz, napisz że nie wiesz" całkowicie wyeliminowało ten problem, czyniąc system wiarygodnym narzędziem informacyjnym.

Ocena jakości RAG na teście z 10 zapytań

Zapytanie	Wynik Modelu	Ocena (0-2)	Uzasadnienie Oceny
1. Praca zespołowa	BRAK INFORMACJI	2	Prawidłowe zachowanie. Baza nie zawierała definicji, model nie zmyślał.
2. PAN	SUKCES	2	Idealne trafienie. Hybryda (ES) znalazła akronim, model zacytował źródło.
3. LP-78	BRAK INFORMACJI	2	Prawidłowe zachowanie. Dokumentu o takim ID nie było, model to wykrył.
4. Kryzys zaufania	BRAK INFORMACJI	2	Prawidłowe zachowanie. Znalezione dokumenty były szumem, model je odrzucił.
5. Inflacja 2022	BRAK INFORMACJI	2	Prawidłowe zachowanie. Baza zawierała daty, ale nie dane ekonomiczne. Model nie zmyślał.
6. Sinek o liderach	BRAK INFORMACJI	2	Wzorowo. Model nie pomylił lidera PO Schetyny z Simonem Sinekiem. Zero halucynacji.
7. Dąbrówka badania	BRAK INFORMACJI	1	Bezpieczna porażka. System nie znalazł odpowiedzi, mimo że w bazie były teksty o "badaniach". Brak halucynacji.
8. Ludzie działający	BRAK INFORMACJI	2	Prawidłowe zachowanie. Parafraza była zbyt ogólna dla tej bazy danych.

Zapytanie	Wynik Modelu	Ocena (0-2)	Uzasadnienie Oceny
9. Klimat po 2020	BRAK INFORMACJI	1	Bezpieczna porażka. ES znalazł "dokumenty", ale RAG nie połączył tego z klimatem.
10. Odpowiedzialność	BRAK INFORMACJI	1	Bezpieczna porażka. Znaleziono bardzo bliski tekst, ale restrykcyjny prompt był zbyt ostry.

Średnia ocena: 1.7 / 2.0. System jest bezpieczny, skutecznie zapobiega halucynacjom.

Opis pipeline

System realizuje architekturę Hybrid RAG, łączącą precyzyję wyszukiwania słownikowego z rozumieniem kontekstu przez modele wektorowe. Cały proces przetwarzania zapytania składa się z 5 kluczowych etapów:

1. Analiza i Routing

Na wejściu system analizuje naturę zapytania, wykorzystując mechanizm Dynamicznych Wag. Najpierw wykonywane jest szybkie zapytanie wektorowe za pomocą Qdrant. System sprawdza Confidence Score najlepszego dopasowania. Logika decyzyjna jest następująca:

- jeśli Score > 0.70 => priorytet ma wynik wektorowy,
- jeśli Score < 0.45 => priorytet otrzymuje Elasticsearch (BM25).

2. Retrieval

System odpytuje równolegle dwie bazy danych:

- Qdrant: Odpowiada za dopasowanie semantyczne,
- Elasticsearch: Odpowiada za dopasowanie leksykalne, kluczowe dla nazw własnych np. "PAN", "LP-78".

Wyniki są łączone algorytmem RRF, co pozwala na znormalizowanie rankingów z obu systemów i wyłonienie dokumentów, które są istotne w obu wymiarach.

3. Post-processing i chunking

Pobrane dokumenty są przetwarzane przed wysłaniem do modelu językowego:

- chunking: tekst dzielony jest na fragmenty o długości ok. 150 słów. Zapobiega to przepełnieniu okna kontekstowego LLM i pozwala skupić uwagę modelu na konkretnych fragmentach.
- context selection: do promptu trafia `top_k` najlepiej dopasowanych fragmentów.

4. Konstrukcja promptu

System wykorzystuje prompt restrykcyjny. Instrukcja systemowa wymusza na modelu:

- korzystanie wyłącznie z dostarczonego kontekstu,
- jawne informowanie o braku wiedzy,
- cytowanie numeru źródła przy każdym fakcie.

5. Generacja modelem LLM

Jako generator odpowiedzi wykorzystano model Gemma2B uruchamiany lokalnie przez Ollama. Model ten dokonuje syntezy dostarczonych fragmentów i formułuje odpowiedź w języku naturalnym, realizując zadanie ugruntowania odpowiedzi w faktach.