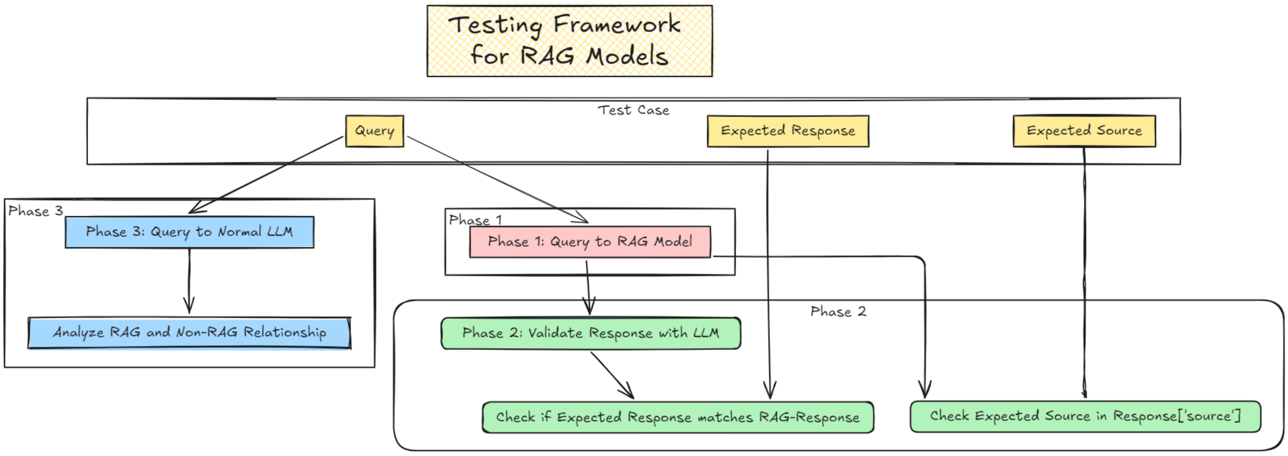
9. Automated Testing

Automated testing was chosen to address the challenges of manual testing, which became increasingly inefficient and error prone as the complexity of the RAG model grew. Automated testing provides significant advantages, such as enhanced consistency, reliability, and scalability, in addition to faster execution and the ability to handle repetitive tasks (***Kumar, 2016***).

Currently, there is no standardized methodology for evaluating RAG systems, but several frameworks have emerged. One of these is the **RAGAS** framework, which introduces two primary evaluation strategies: **Faithfulness** and **Answer Relevance**. **Faithfulness** assesses whether the answer is grounded in the provided context, while **Answer Relevance** determines if the answer properly addresses the question. We selected a test set consisting of fifty questions, as this provides a solid foundation for evaluating model performance. Each question in the set includes an *expected answer* and *expected source*, creating a standardized way of testing. These fifty questions are derived from various scientific papers on NEET that we have already ingested into our database.

Given our limited GPU resources, we developed a testing framework that optimizes the process by dividing it into three phases using *subprocesses*, allowing us to conduct detailed testing without overloading system resources. This framework is built upon the principles of **Faithfulness** and **Answer Relevance** as outlined in the **RAGAS** methodology, ensuring that responses both contain the correct context and directly address the questions posed.


Testing Framework for RAG Models
Test Case
Query
Expected Response
Expected Source
Phase 3
Phase 3: Query to Normal LLM
Analyze RAG and Non-RAG Relationship
Phase 1
Phase 1: Query to RAG Model
Phase 2
Phase 2: Validate Response with LLM
Check if Expected Response matches RAG-Response
Check Expected Source in Response['source']

Phase 1 involves querying the RAG model with questions taken directly from the test set. The model generates an answer along with a source, which is logged for further evaluation. **Phase 2** evaluates the answers produced in **Phase 1** by prompting *Normistral-7B* to compare these answers to the expected answers from the test set. We also perform a source verification in this phase by checking if the expected source is present within the sources returned by the RAG model's response. If both the answer and source match the expected results, the test is marked as passed; otherwise, it is considered failed. Additionally, we included both positive and negative test cases to assess *Normistral-7B*'s ability to identify mismatches and validate accuracy.

Phase 3 repeats the query process from **Phase 1** but with the RAG model disabled. This phase allows us to directly evaluate *Normistral-7B's* standalone responses, comparing them to the RAG-enhanced responses from **Phase 1**. By examining the differences between these responses, we gain insights into how the RAG model affects answer accuracy and relevance.

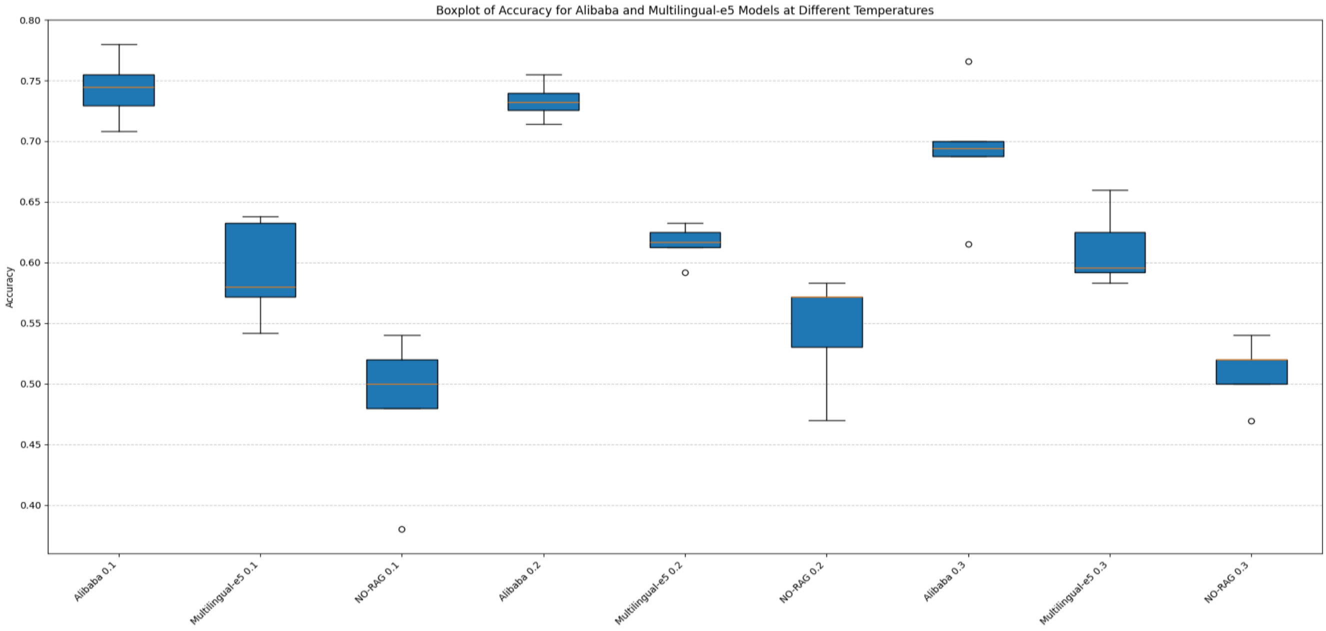
Furthermore, we measured the average response generation time for both RAG-enabled and non-RAG setups to analyse performance in terms of speed. This provides valuable insights into the trade-offs between RAG's enhanced accuracy and the potential increase in latency.

This testing framework enabled us to compare the RAG model's performance with alternative embedding models, including Alibaba and Multilingual-E5, as well as *Normistral-7B* alone (NO-RAG). We varied the temperature settings between 0.1, 0.2, and 0.3 to understand how these modifications affect overall accuracy.

10. Results

This section presents the accuracy, response quality, and efficiency of different models assessed within our framework. The results include a comparative analysis of embedding models and non-RAG setups, an examination of RAG vs. NO-RAG responses, and an assessment of response times for each approach.

10.1 Model Accuracy Comparison

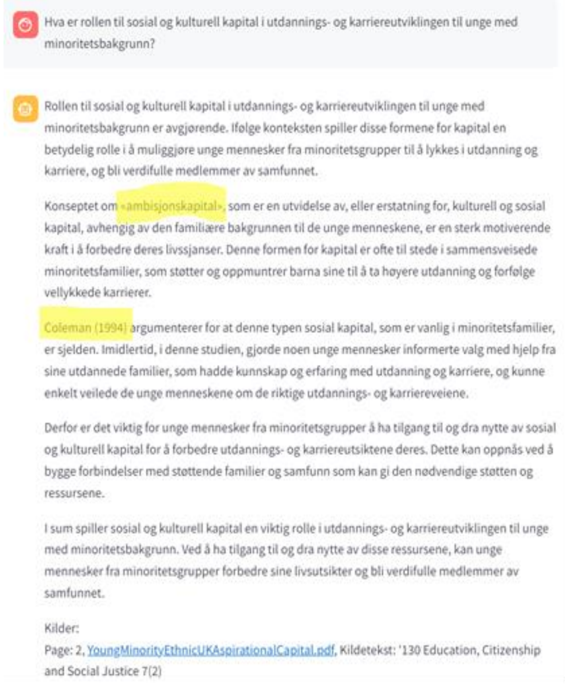
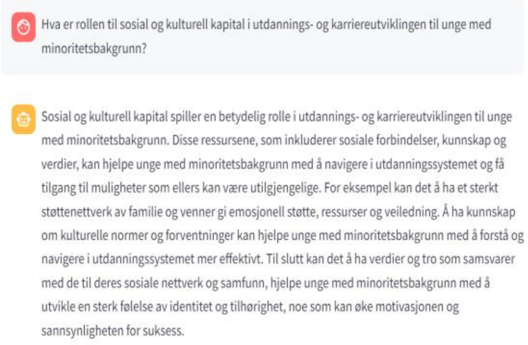


We evaluated the accuracy of different embedding models (Alibaba and Multilingual-E5) and the NO-RAG setup using various temperature settings (0.1, 0.2, and 0.3). The results highlight a clear difference in performance across the models and temperature levels.

- **Alibaba** consistently performed the best across all temperature settings, achieving 75% accuracy at 0.1, slightly lower at 72-73% at 0.2, and 69% at 0.3. This suggests that Alibaba's accuracy decreases slightly as the temperature increases, probably due to increased response flexibility.
- **Multilingual-E5** showed moderate performance, reaching 58% accuracy at 0.1, peaking at 61% at 0.2, and dropping slightly to 59% at 0.3. These results indicate that Multilingual-E5 performs most accurately at a mid-range temperature setting (0.2).
- **NO-RAG** (Normistral-7B without retrieval augmentation) showed the lowest accuracy across the board, with 50% at 0.1, increasing slightly to 55% at 0.2, but dropping significantly to 0.5% at 0.3. This steep decline at 0.3 suggests that NO-RAG struggles to maintain accuracy as temperature increases, likely due to its lack of grounding in external sources.

These findings indicate that the RAG setup significantly enhances accuracy, with Alibaba proving the most dependable among the tested embedding models, especially at lower temperature settings.

10.2 RAG vs. NO-RAG Responses

RAG Response	NO-RAG Response
 Hva er rollen til sosial og kulturell kapital i utdannings- og karriereutviklingen til unge med minoritetsbakgrunn? Rollen til sosial og kulturell kapital i utdannings- og karriereutviklingen til unge med minoritetsbakgrunn er avgjørende. Ifølge konteksten spiller disse formene for kapital en betydelig rolle i å muliggjøre unge mennesker fra minoritetsgrupper til å lykkes i utdanning og karriere, og bli verdifulle medlemmer av samfunnet. Konseptet om «ambisjonskapital», som er en utvidelse av, eller erstatning for, kulturell og sosial kapital, avhengig av den familiære bakgrunnen til de unge menneskene, er en sterk motiverende kraft i å forbedre deres livssjanser. Denne formen for kapital er ofte til stede i sammensveisede minoritetsfamilier, som støtter og oppmuntrer barna sine til å ta høyere utdanning og forfølge vellykkede karrierer. Coleman (1994) argumenterer for at denne typen sosial kapital, som er vanlig i minoritetsfamilier, er sjelden. Imidlertid, i denne studien, gjorde noen unge mennesker informerte valg med hjelp fra sine utdannede familier, som hadde kunnskap og erfaring med utdanning og karriere, og kunne enkelt veilede de unge menneskene om de riktige utdannings- og karriereveiene. Derfor er det viktig for unge mennesker fra minoritetsgrupper å ha tilgang til og dra nytte av sosial og kulturell kapital for å forbedre utdannings- og karriereutsiktene deres. Dette kan oppnås ved å bygge forbindelser med støttende familier og samfunn som kan gi den nødvendige støtten og ressursene. I sum spiller sosial og kulturell kapital en viktig rolle i utdannings- og karriereutviklingen til unge med minoritetsbakgrunn. Ved å ha tilgang til og dra nytte av disse ressursene, kan unge mennesker fra minoritetsgrupper forbedre sine livsutsikter og bli verdifulle medlemmer av samfunnet. Kilder: Page: 2, YoungMinorityEthnicUKAspirationalCapital.pdf, Kildetekst: '130 Education, Citizenship and Social Justice 7(2)	 Hva er rollen til sosial og kulturell kapital i utdannings- og karriereutviklingen til unge med minoritetsbakgrunn? Sosial og kulturell kapital spiller en betydelig rolle i utdannings- og karriereutviklingen til unge med minoritetsbakgrunn. Disse ressursene, som inkluderer sosiale forbindelser, kunnskap og verdier, kan hjelpe unge med minoritetsbakgrunn med å navigere i utdanningssystemet og få tilgang til muligheter som ellers kan være utilgjengelige. For eksempel kan det å ha et sterkt støttenettverk av familie og venner gi emosjonell støtte, ressurser og veiledning. Å ha kunnskap om kulturelle normer og forventninger kan hjelpe unge med minoritetsbakgrunn med å forstå og navigere i utdanningssystemet mer effektivt. Til slutt kan det å ha verdier og tro som samsvarer med de til deres sosiale nettverk og samfunn, hjelpe unge med minoritetsbakgrunn med å utvikle en sterk følelse av identitet og tilhørighet, noe som kan øke motivasjonen og sannsynligheten for suksess.

In comparing the content of RAG-generated responses to those from the NO-RAG setup, notable differences in response quality and reliability emerged:

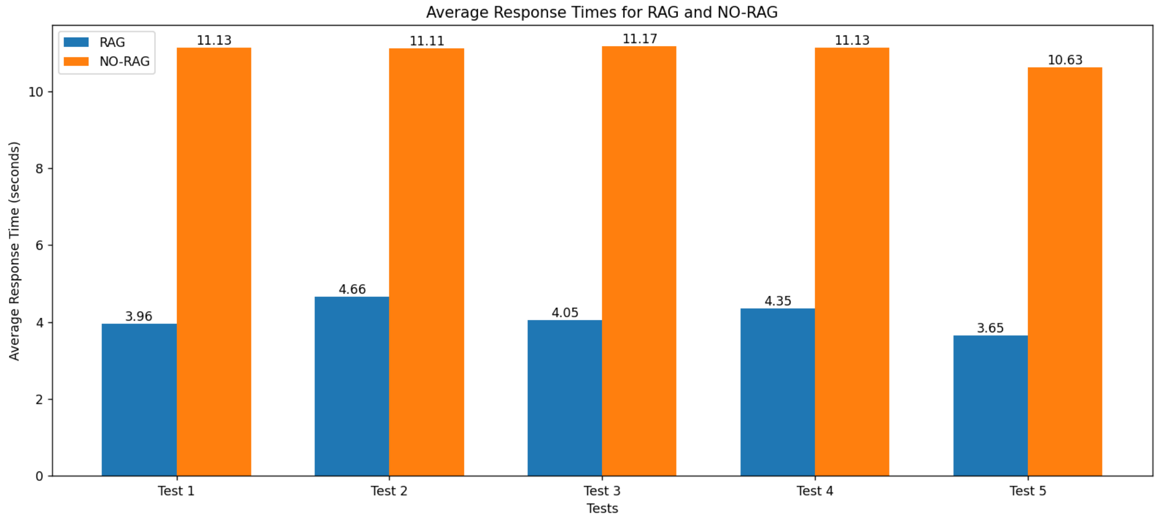
RAG Responses: Responses generated by the RAG model consistently utilized sources and specific figures to support arguments. The inclusion of sources allowed users to verify the information directly, ensuring a high degree of transparency and credibility in RAG responses.

NO-RAG Responses: Although NO-RAG occasionally included sources, there was no reliable way to verify the accuracy of these sources. As NO-RAG lacks the retrieval component, the cited sources could not be confirmed as accurate or relevant to the question, leading to a potential lack of trustworthiness in its answers.

This comparison underscores the advantage of RAG's retrieval mechanism, as it provides grounded, verifiable information, whereas NO-RAG may generate answers with unsupported or unverifiable references.

10.3 Response Time Analysis: RAG vs. NO-RAG

The response times between the RAG and NO-RAG setups differed significantly. We conducted ten tests, each containing fifty questions, and recorded the average response time per test for both setups:



The mean response time for RAG was 4.1342 seconds, while the mean response time for NO-RAG was significantly higher, at 11.034 seconds. This data indicates that RAG is, on average, *2.67* times faster (or *166.9%* faster) than NO-RAG in generating responses.

The speed advantage of the RAG model is likely due to its ability to retrieve contextually relevant information directly, whereas NO-RAG requires longer generation times without external grounding, resulting in a slower overall response.

10.4 Limitations of the Testing Framework

While our testing framework provided valuable insights into model performance, it also has certain limitations that should be considered:

Model Size: The evaluations were conducted using a small model (Normistral-7B), which may limit the model's ability to fully comprehend complex questions and generate highly accurate responses, particularly when managing scientific content.

Uncertain Test Cases: The binary true/false setup in our testing framework occasionally resulted in ambiguous cases, as Normistral-7B sometimes struggled to clearly return "true" or "false." These uncertain cases introduce a degree of inaccuracy in our accuracy assessments.

Matching Answers with Incorrect Sources: In some cases, models generated accurate answers but provided incorrect sources. This may occur because multiple documents contain similar information, making it difficult for the model to pinpoint the exact source. As a result, the model may appear less accurate in source verification, even if the answer is correct.