LLM Evaluation Report

Executive Summary

Methodology

The evaluation methodology for the WAISE project is designed to assess self-hosted Large Language Models (LLMs) within XWiki, focusing on their performance in knowledge management and technical support tasks.

- Task Selection: Tasks are chosen to cover essential areas such as content summarization, question answering, and other knowledge management functions.
- Benchmarking Framework: A benchmarking framework is established to measure LLM performance using automated metrics and manual evaluation.
- Ethical Considerations: The evaluation includes considerations for data privacy and the environmental impact of the models.
- Iterative Improvement: The methodology involves continuous refinement of models and evaluation metrics based on feedback.

Find the full methodology document at:
https://design.xwiki.org/xwiki/bin/view/Proposal/X-AI/WAISE/Evaluation%20Methodology/

Metrics

Answer Relevancy

Measures the quality of the RAG pipeline's generator by evaluating how relevant the actual output is compared to the provided input. Score is calculated as: Number of Relevant Statements / Total Number of Statements.

Faithfulness

Evaluates whether the actual output factually aligns with the contents of the retrieval context. Score is calculated as: Number of Truthful Claims / Total Number of Claims.

Contextual Precision

Measures the RAG pipeline's retriever by evaluating whether relevant nodes in the retrieval context are ranked higher than irrelevant ones.

Contextual Recall

Evaluates the extent to which the retrieval context aligns with the expected output.

Custom Context Relevancy

Assesses how well the retrieval context aligns with the information required to generate the expected output. It evaluates the proportion of information from the expected answer found within the retrieval context.

Note: All metrics are self-explaining LLM-Eval, providing reasons for their scores.

Evaluation Criteria:

Summarization: Alignment, Coverage
Text_generation: score
Rag-qa: AnswerRelevancy, Faithfulness, ContextualPrecision, ContextualRecall, CustomContextRelevancy

Model Information

Model	Context Length	Provider	License
GPT4o	128k	OpenAI	Proprietary
GPT4o-mini	128k	OpenAI	Proprietary
claude3_5_sonet	200k	OpenRouter	Proprietary
mistral2_large	128k	Mistral	Mistral Research License
llama3_1_402b	128k	OpenRouter	LLAMA 3.1
llama3_1_8b_Q4	128k	Ollama	LLAMA 3.1
mixtral-8x22b	16k	Mistral	Apache 2.0
mistral-nemo_12b_Q4	128k	Ollama	Apache 2.0
gemma2_9B_Q4	8k	Ollama	Gemma ToU
phi3_mini-128k_4b_Q4	128k	Ollama	MIT License
phi3_medium-128k_14b_Q4	128k	Ollama	MIT License
command-r_35B_Q4	128k	Ollama	CCA-NonCommercial 4.0
qwen2_7b_Q4	128k	Ollama	Apache 2.0

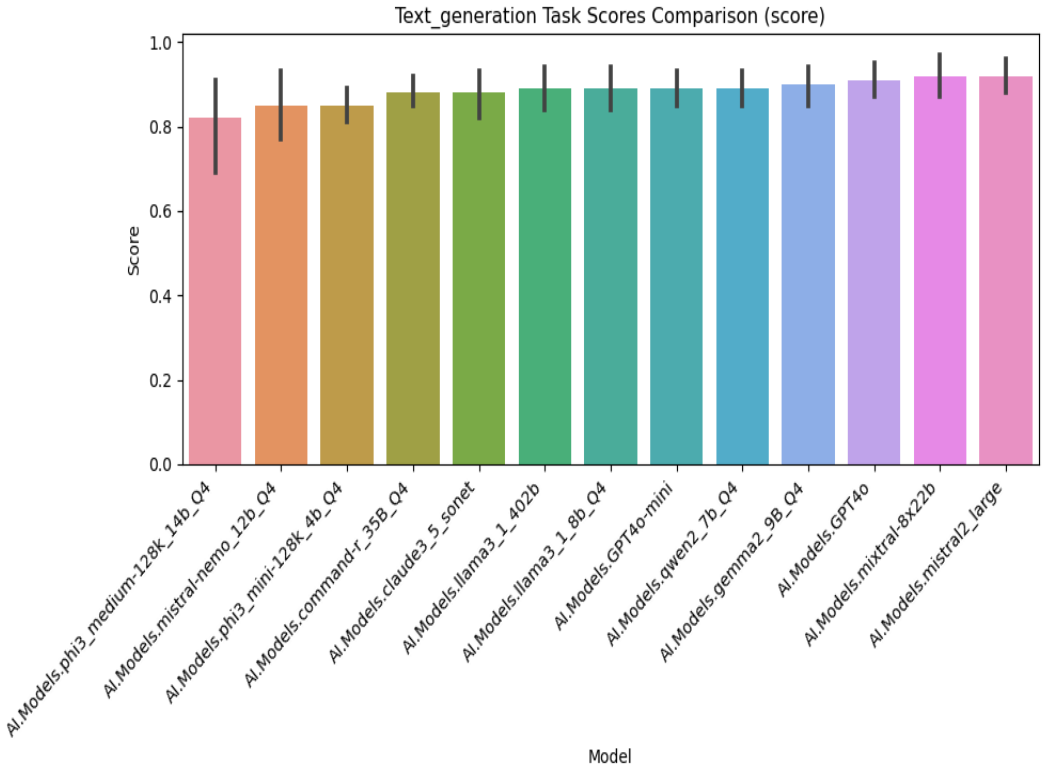
Setup Information

Task	Model	Temperature	Power Measurement

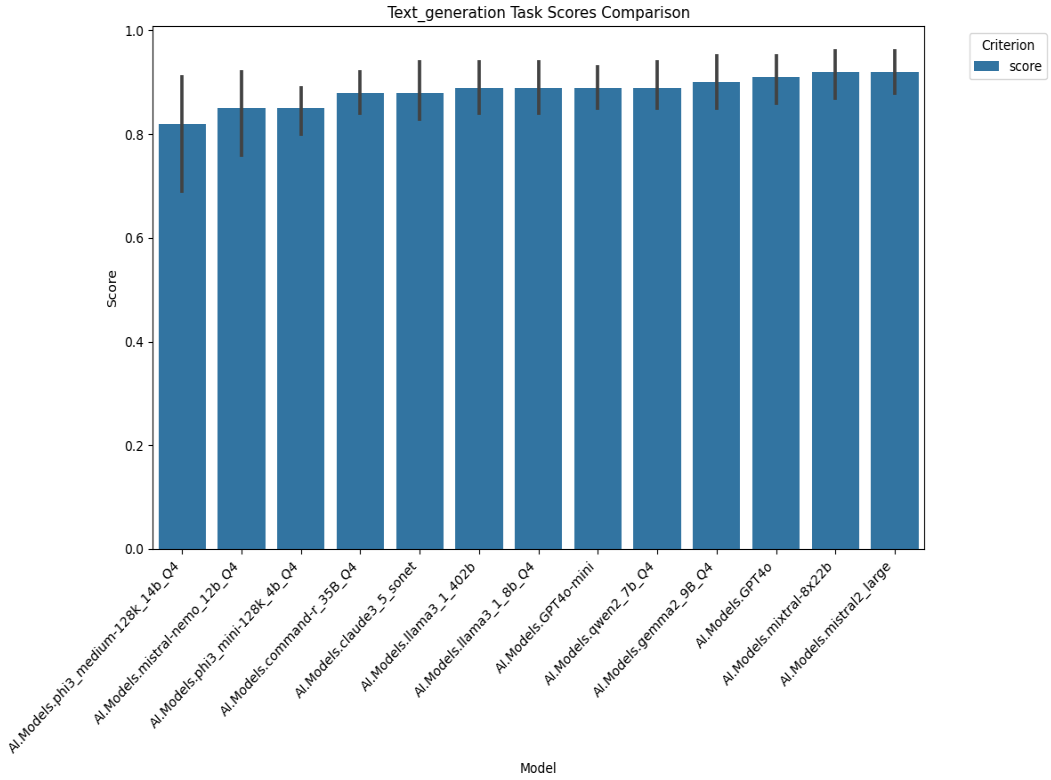
RAG-qa	AI.Models.qa_GPT4o	0.8	No
RAG-qa	AI.Models.qa_GPT4o-mini	0.8	No
RAG-qa	AI.Models.qa_claude3_5_sonet	0.8	No
RAG-qa	AI.Models.qa_mistral2_large	0.8	No
RAG-qa	AI.Models.qa_llama3_1_402b	0.45	No
RAG-qa	AI.Models.qa_llama3_1_8b_Q4	0.45	Yes
RAG-qa	AI.Models.qa_mixtral-8x22b	0.8	No
RAG-qa	AI.Models.qa_mistral-nemo_12b_Q4	0.3	Yes
RAG-qa	AI.Models.qa_gemma2_9B_Q4	0.3	Yes
RAG-qa	AI.Models.qa_phi3_mini-128k_4b_Q4	0	Yes
RAG-qa	AI.Models.qa_phi3_medium-128k_14b_Q4	0.8	Yes
RAG-qa	AI.Models.qa_command-r_35B_Q4	0.3	Yes
RAG-qa	AI.Models.qa_qwen2_7b_Q4	0	Yes
summarization	AI.Models.GPT4o	0.3	No
summarization	AI.Models.GPT4o-mini	0.3	No
summarization	AI.Models.claude3_5_sonet	0.3	No
summarization	AI.Models.mistral2_large	0.3	No
summarization	AI.Models.llama3_1_402b	0.3	No
summarization	AI.Models.llama3_1_8b_Q4	0.3	Yes
summarization	AI.Models.mixtral-8x22b	0.8	No
summarization	AI.Models.mistral-nemo_12b_Q4	0.3	Yes
summarization	AI.Models.gemma2_9B_Q4	0.3	Yes
summarization	AI.Models.phi3_mini-128k_4b_Q4	0.3	Yes
summarization	AI.Models.phi3_medium-128k_14b_Q4	0.8	Yes
summarization	AI.Models.command-r_35B_Q4	0.3	Yes
summarization	AI.Models.qwen2_7b_Q4	0.3	Yes

text_generation	AI.Models.GPT4o	0.3	No
text_generation	AI.Models.GPT4o-mini	0.8	No
text_generation	AI.Models.claude3_5_sonet	0.8	No
text_generation	AI.Models.mistral2_large	0.8	No
text_generation	AI.Models.llama3_1_402b	0.8	No
text_generation	AI.Models.llama3_1_8b_Q4	0.8	Yes
text_generation	AI.Models.mixtral-8x22b	0.8	No
text_generation	AI.Models.mistral-nemo_12b_Q4	0.8	Yes
text_generation	AI.Models.gemma2_9B_Q4	0.8	Yes
text_generation	AI.Models.phi3_mini-128k_4b_Q4	0.8	Yes
text_generation	AI.Models.phi3_medium-128k_14b_Q4	0.8	Yes
text_generation	AI.Models.command-r_35B_Q4	0.8	Yes
text_generation	AI.Models.qwen2_7b_Q4	0.8	Yes

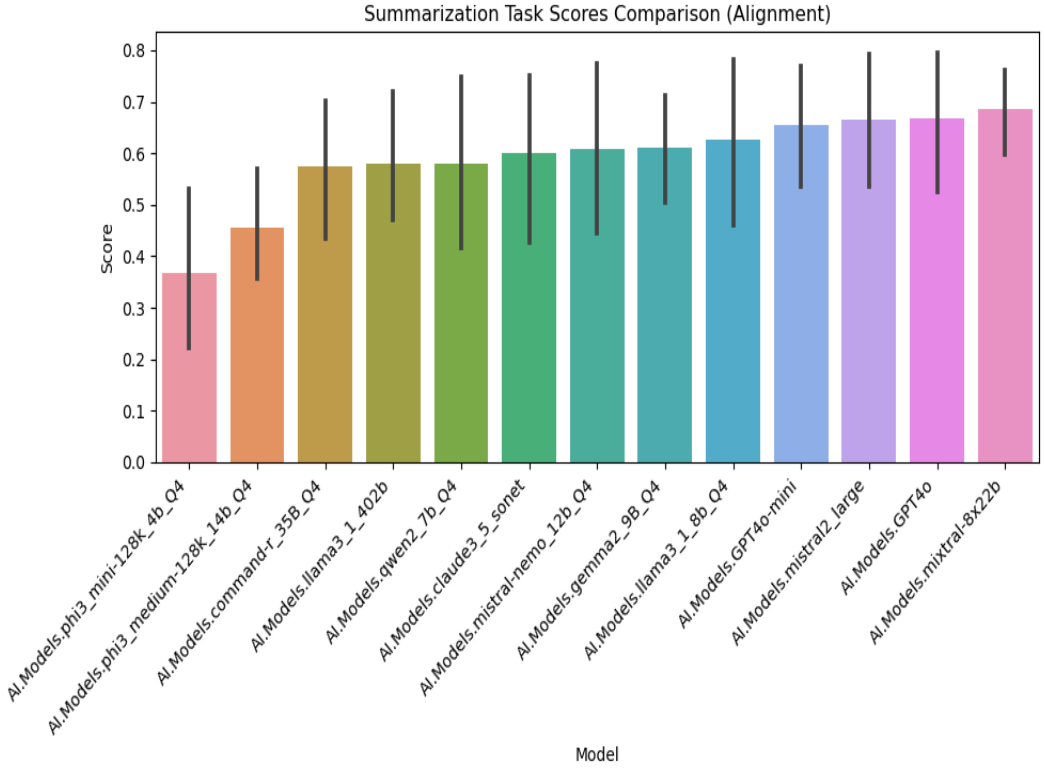
Text_generation Task Results

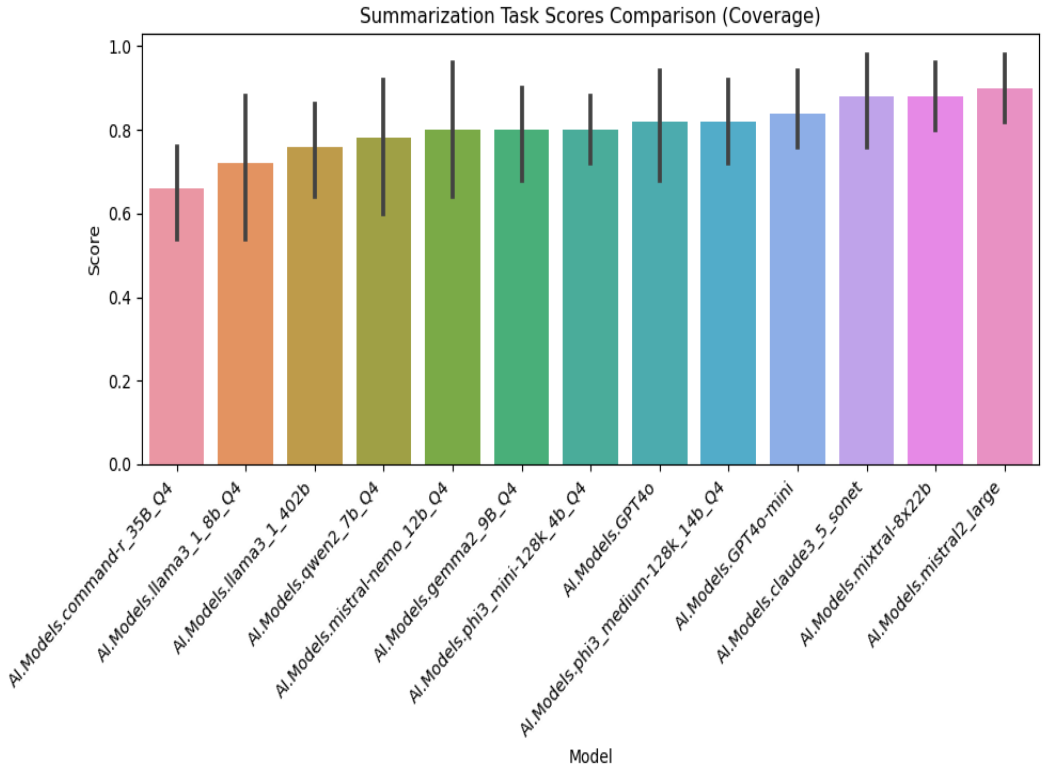


Grouped Results

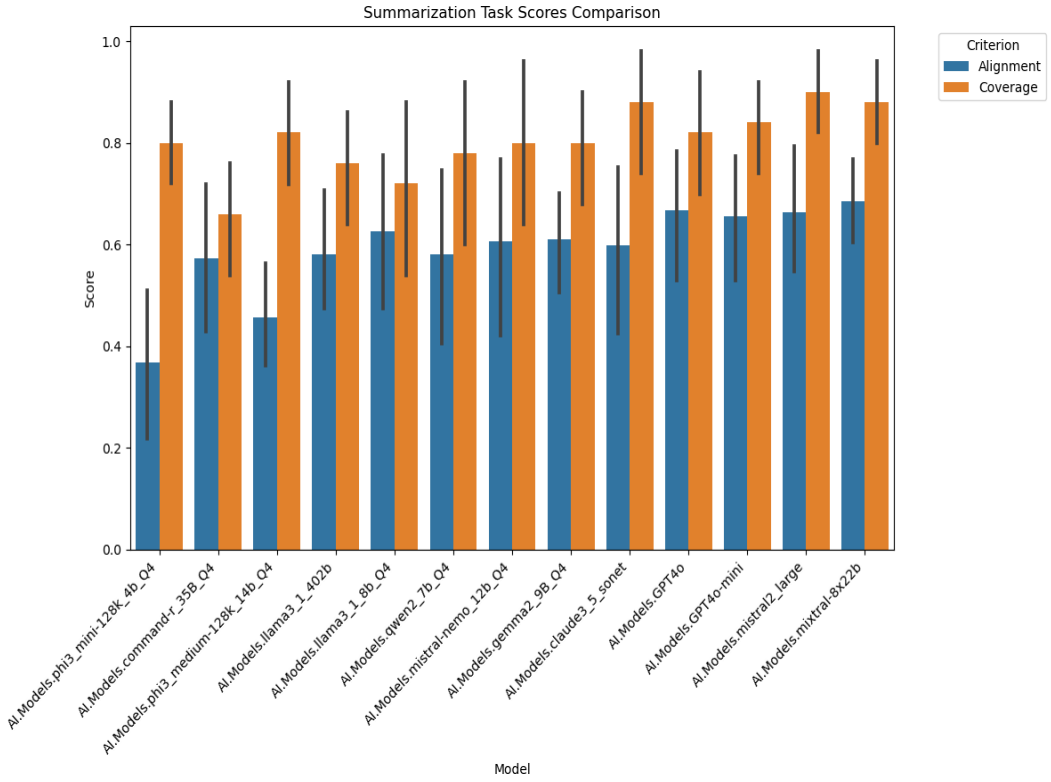


Summarization Task Results

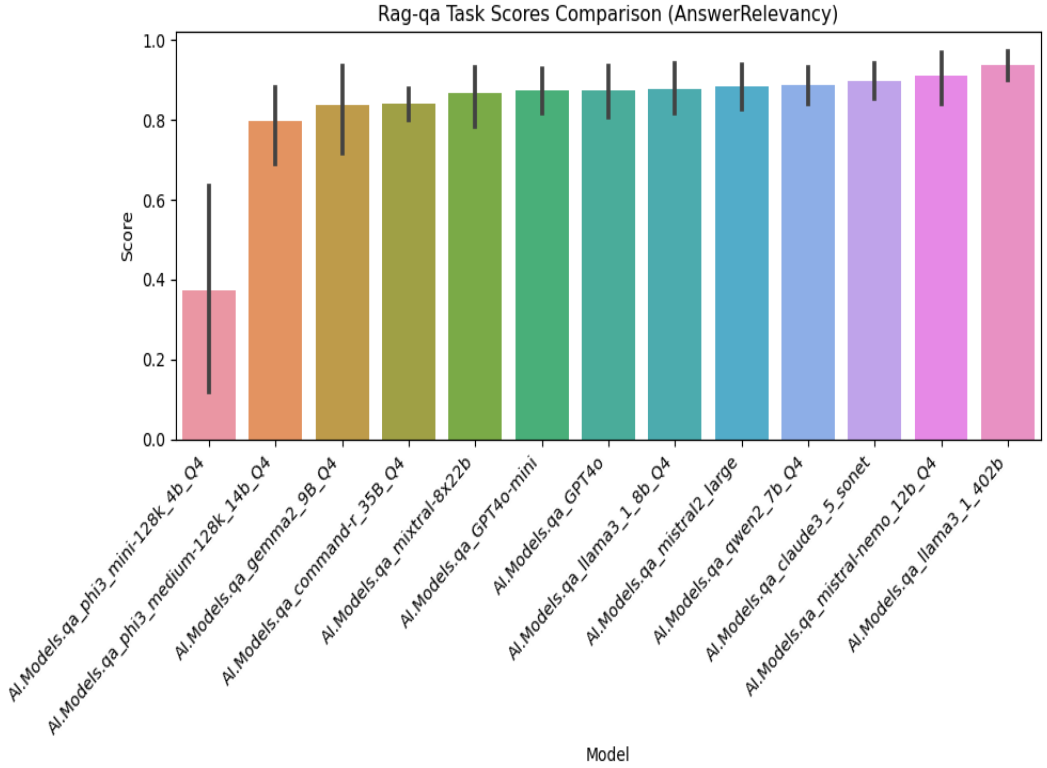

Summarization Task Scores Comparison (Alignment)
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
AI.Models.phi3_mini-128k_4b_Q4
AI.Models.phi3_medium-128k_14b_Q4
AI.Models.command-r_35B_Q4
AI.Models.llama3_1_402b
AI.Models.qwen2_7b_Q4
AI.Models.claude3_5_sonet
AI.Models.mistral-nemo_12b_Q4
AI.Models.gemma2_9B_Q4
AI.Models.llama3_1_8b_Q4
AI.Models.GPT4o-mini
AI.Models.mistral2_large
AI.Models.GPT4o
AI.Models.mixtral-8x22b
Model

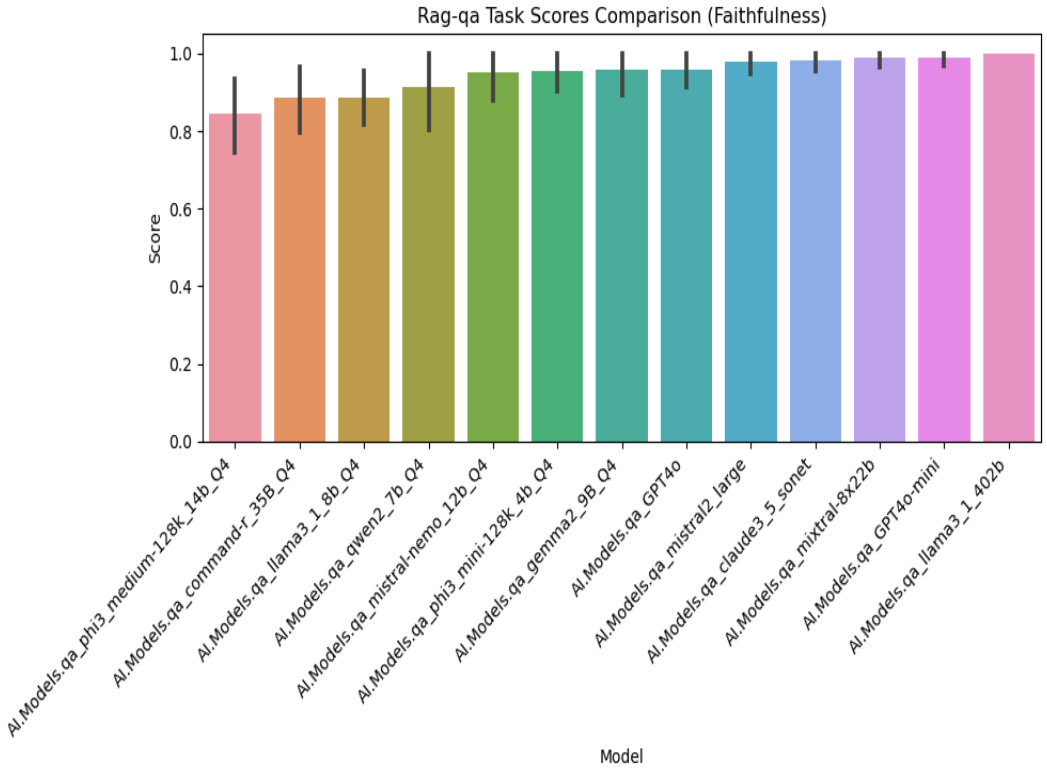

Summarization Task Scores Comparison (Coverage)
Score
0.0
0.2
0.4
0.6
0.8
1.0
AI.Models.command-r_35B_Q4
AI.Models.llama3_1_8b_Q4
AI.Models.llama3_1_402b
AI.Models.qwen2_7b_Q4
AI.Models.mistral-nemo_12b_Q4
AI.Models.gemma2_9B_Q4
AI.Models.phi3_mini-128k_4b_Q4
AI.Models.GPT4o
AI.Models.phi3_medium-128k_14b_Q4
AI.Models.GPT4o-mini
AI.Models.claude3_5_sonet
AI.Models.mixtral-8x22b
AI.Models.mistral2_large
Model

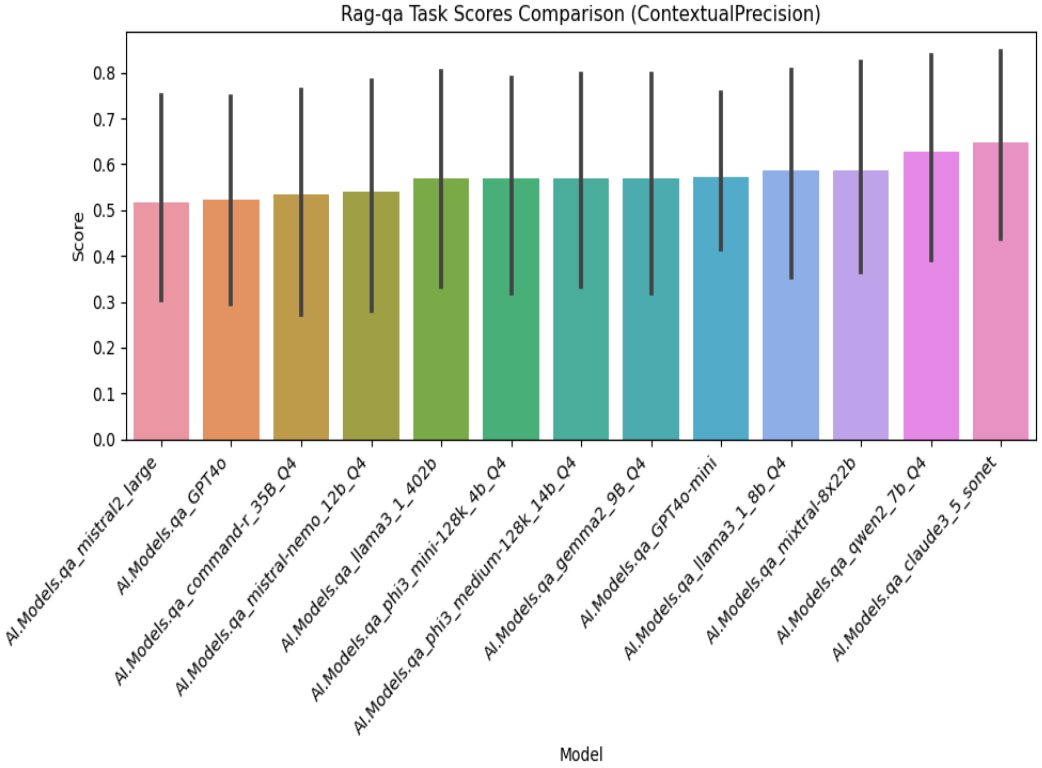
Grouped Results

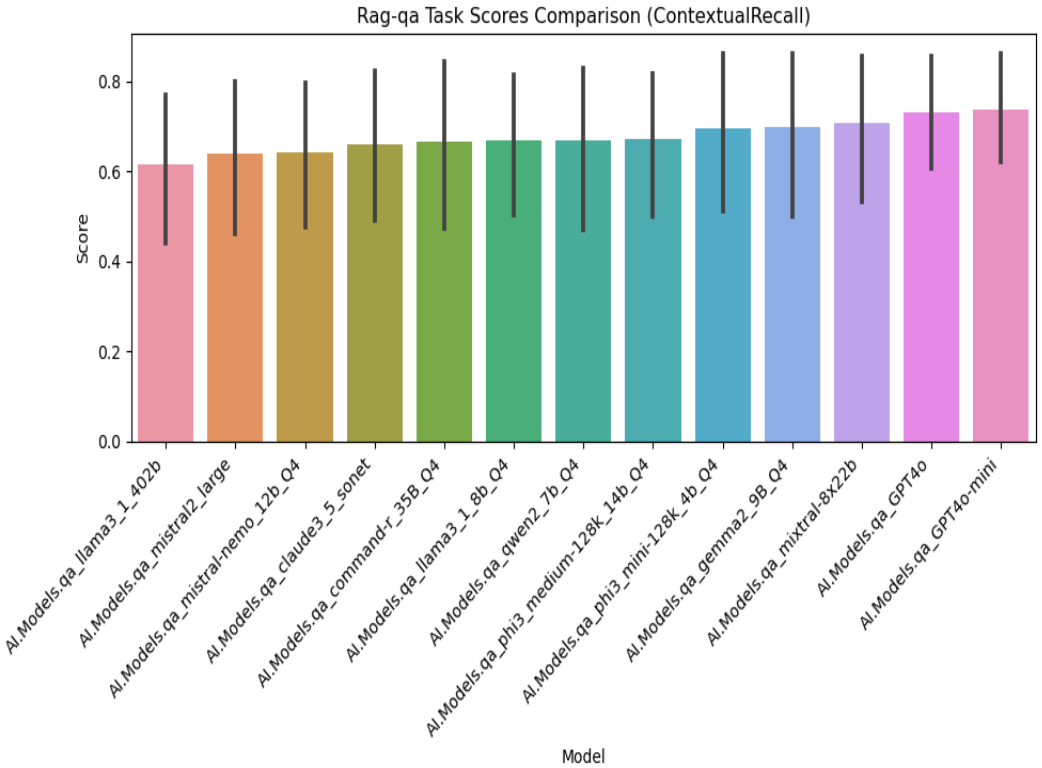


Rag-qa Task Results

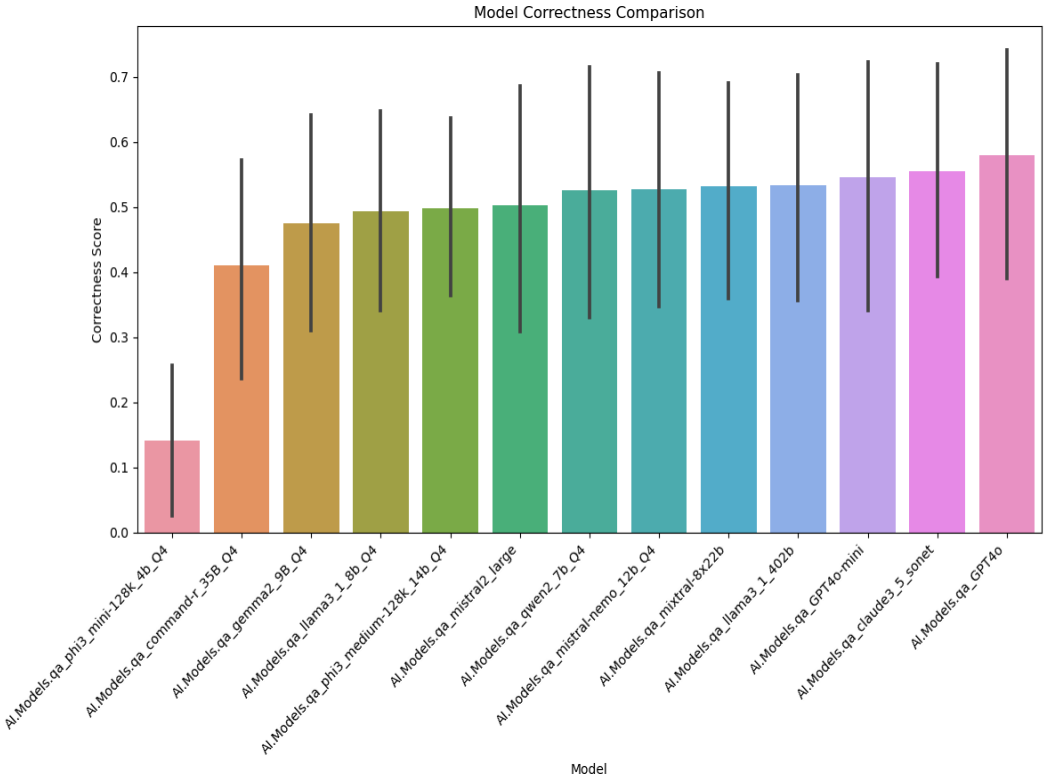

Rag-qa Task Scores Comparison (AnswerRelevancy)
Score
1.0
0.8
0.6
0.4
0.2
0.0
AI.Models.qa_phi3_mini-128k_4b_Q4
AI.Models.qa_phi3_medium-128k_14b_Q4
AI.Models.qa_gemma2_9B_Q4
AI.Models.qa_command-r_35B_Q4
AI.Models.qa_mixtral-8x22b
AI.Models.qa_GPT4o-mini
AI.Models.qa_GPT4o
AI.Models.qa_llama3_1_8b_Q4
AI.Models.qa_mistral2_large
AI.Models.qa_qwen2_7b_Q4
AI.Models.qa_claude3_5_sonet
AI.Models.qa_mistral-nemo_12b_Q4
AI.Models.qa_llama3_1_402b
Model


Rag-qa Task Scores Comparison (Faithfulness)
Score
1.0
0.8
0.6
0.4
0.2
0.0
AI.Models.qa_phi3_medium-128k_14b_Q4
AI.Models.qa_command-r_35B_Q4
AI.Models.qa_llama3_1_8b_Q4
AI.Models.qa_qwen2_7b_Q4
AI.Models.qa_mistral-nemo_12b_Q4
AI.Models.qa_phi3_mini-128k_4b_Q4
AI.Models.qa_gemma2_9B_Q4
AI.Models.qa_GPT4o
AI.Models.qa_mistral2_large
AI.Models.qa_claude3_5_sonet
AI.Models.qa_mixtral-8x22b
AI.Models.qa_GPT4o-mini
AI.Models.qa_llama3_1_402b
Model

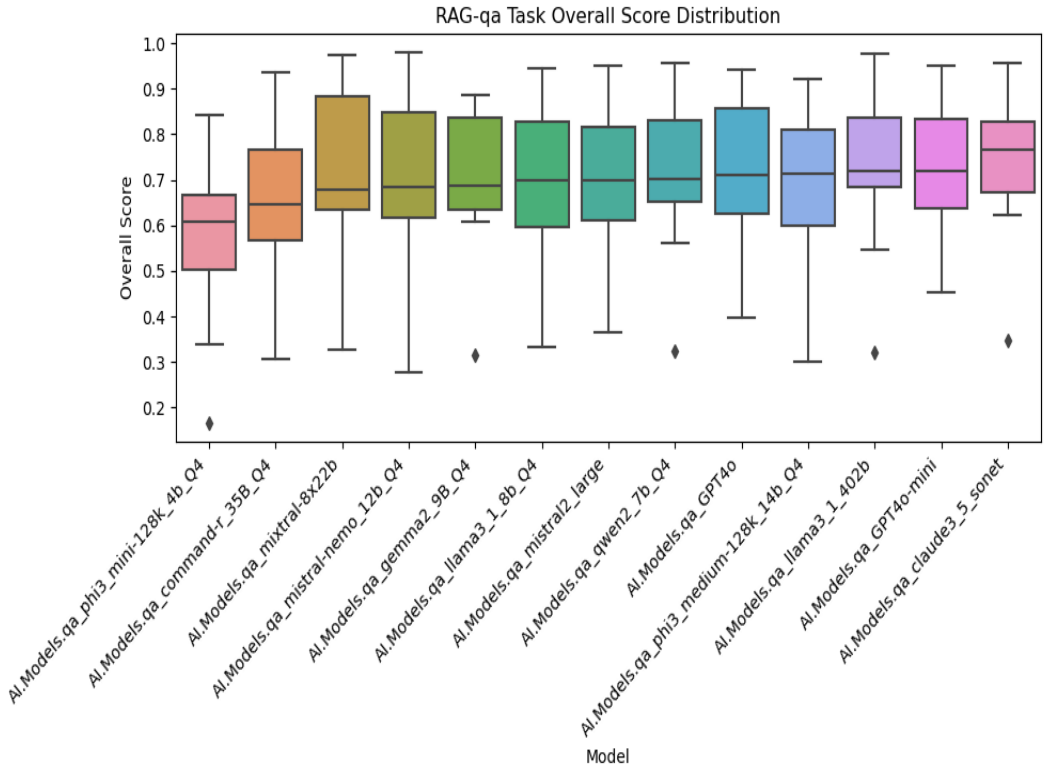

Rag-qa Task Scores Comparison (ContextualPrecision)
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
AI.Models.qa_mistral2_large
AI.Models.qa_GPT4o
AI.Models.qa_command-r_35B_Q4
AI.Models.qa_mistral-nemo_12b_Q4
AI.Models.qa_llama3_1_402b
AI.Models.qa_phi3_mini-128k_4b_Q4
AI.Models.qa_phi3_medium-128k_14b_Q4
AI.Models.qa_gemma2_9B_Q4
AI.Models.qa_GPT4o-mini
AI.Models.qa_llama3_1_8b_Q4
AI.Models.qa_mixtral-8x22b
AI.Models.qa_qwen2_7b_Q4
AI.Models.qa_claude3_5_sonet
Model


Rag-qa Task Scores Comparison (ContextualRecall)
Score
0.0
0.2
0.4
0.6
0.8
AI.Models.qa_llama3_1_402b
AI.Models.qa_mistral2_large
AI.Models.qa_mistral-nemo_12b_Q4
AI.Models.qa_claude3_5_sonet
AI.Models.qa_command-r_35B_Q4
AI.Models.qa_llama3_1_8b_Q4
AI.Models.qa_qwen2_7b_Q4
AI.Models.qa_phi3_medium-128k_14b_Q4
AI.Models.qa_phi3_mini-128k_4b_Q4
AI.Models.qa_gemma2_9B_Q4
AI.Models.qa_mixtral-8x22b
AI.Models.qa_GPT4o
AI.Models.qa_GPT4o-mini
Model

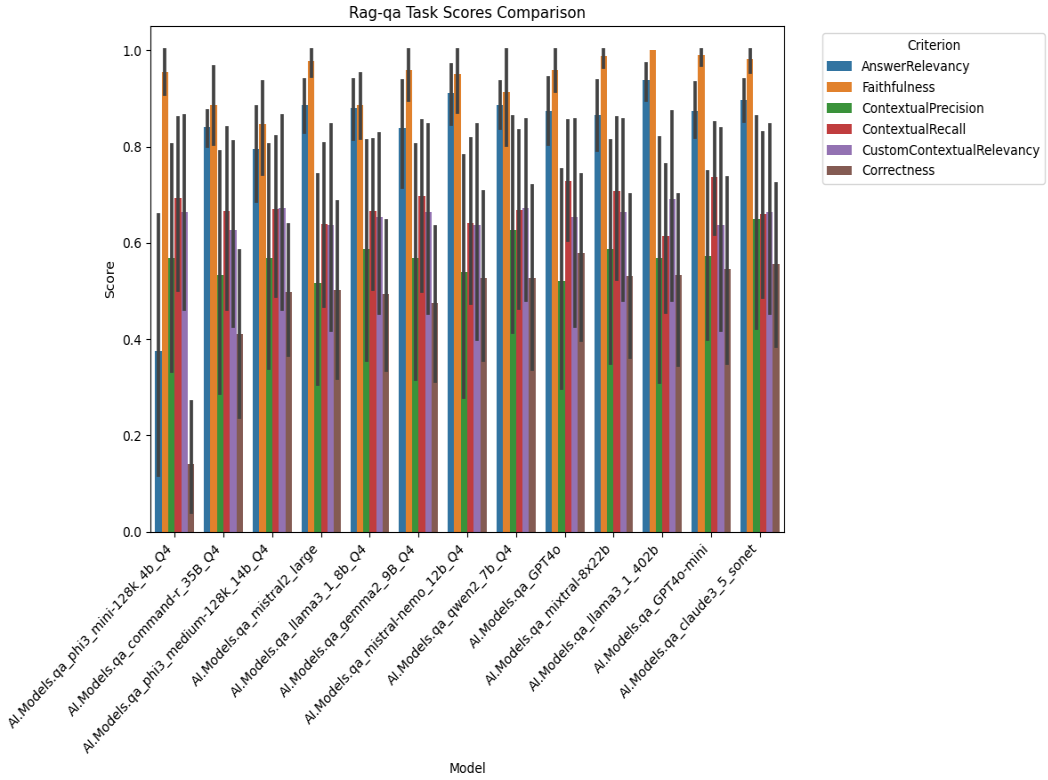
Model Correctness Comparison



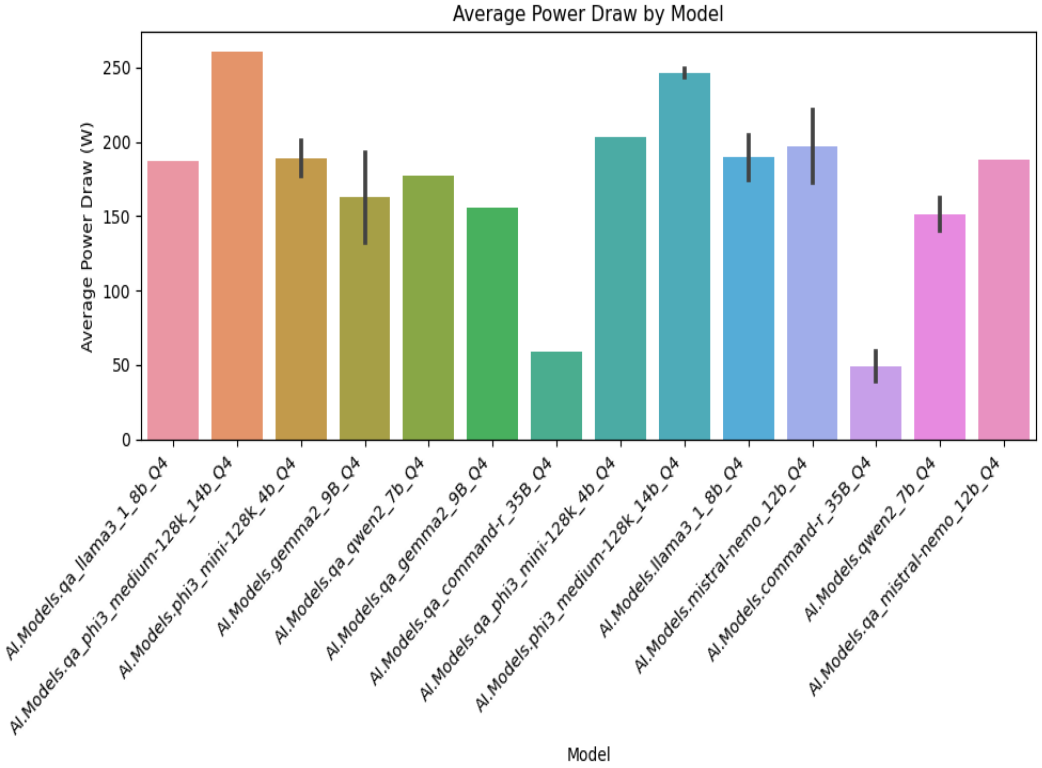
Overall Score Distribution

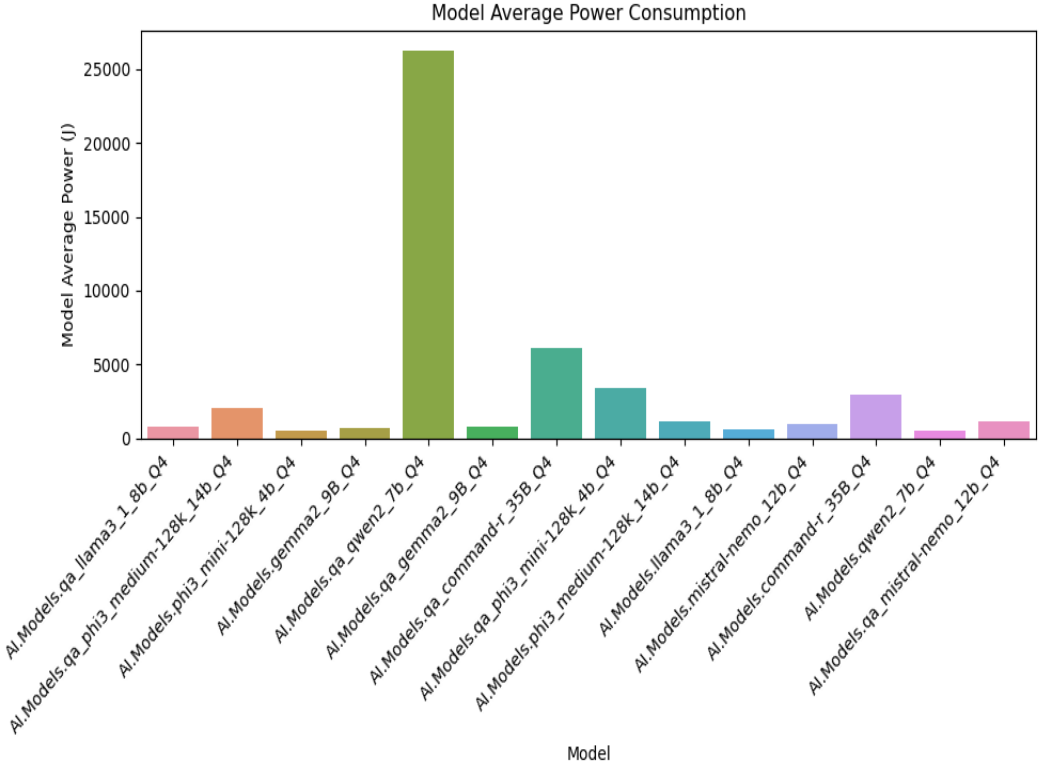


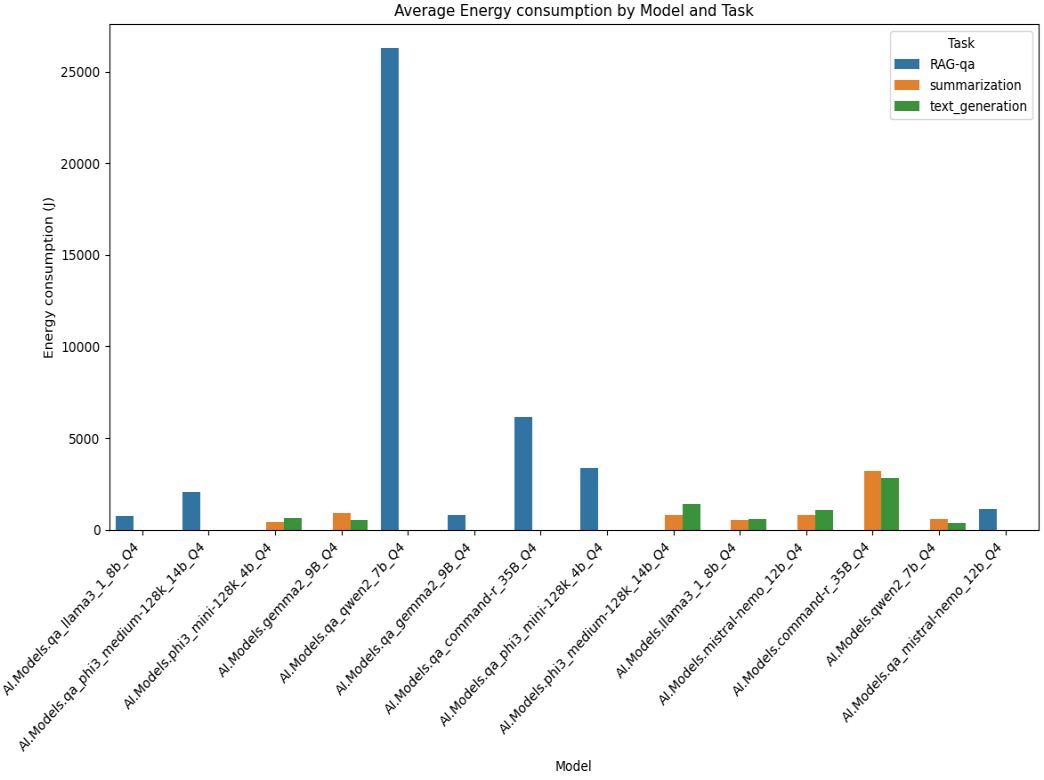
Grouped Results

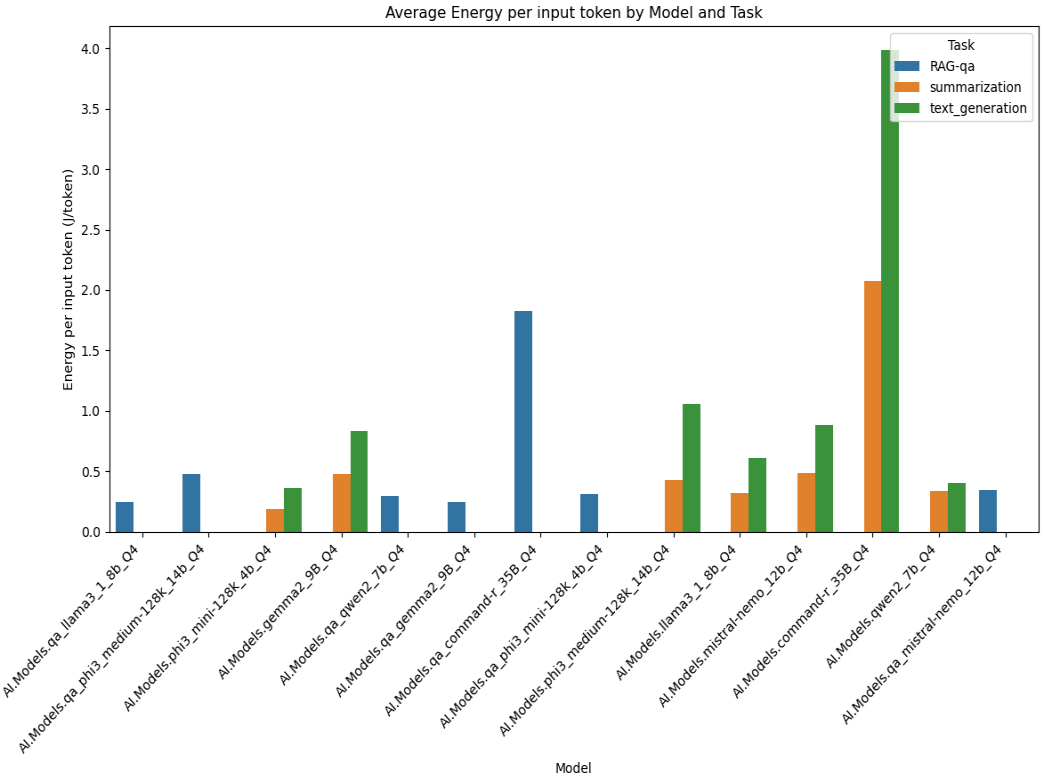


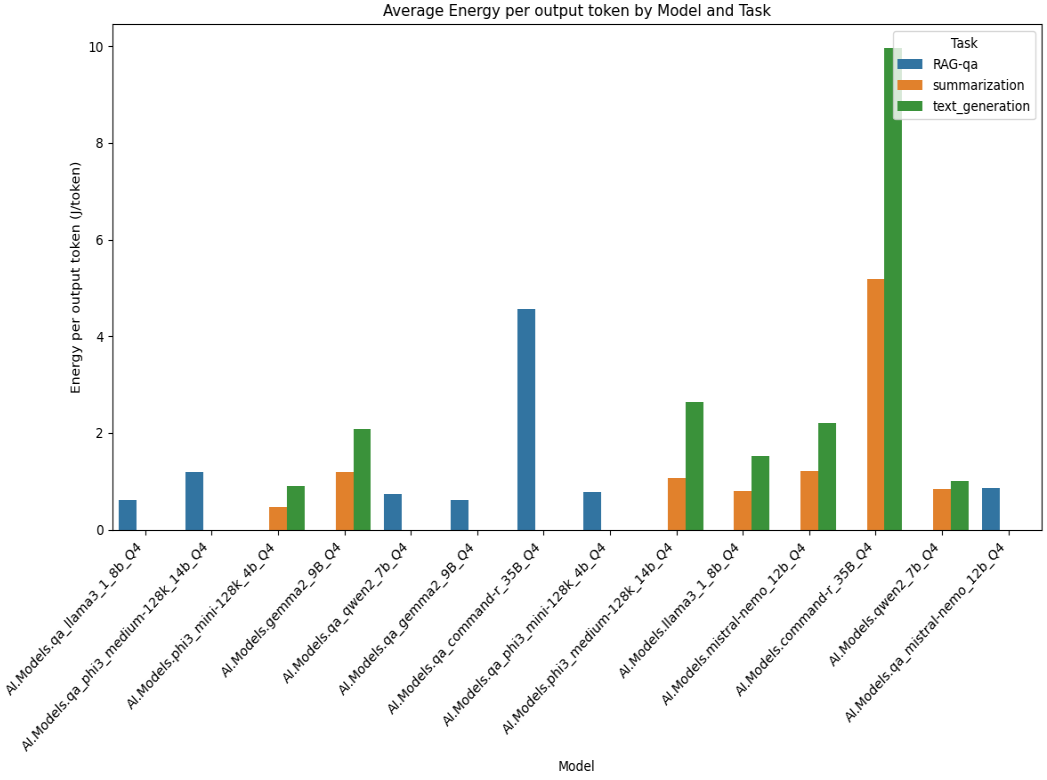
Power Consumption Results

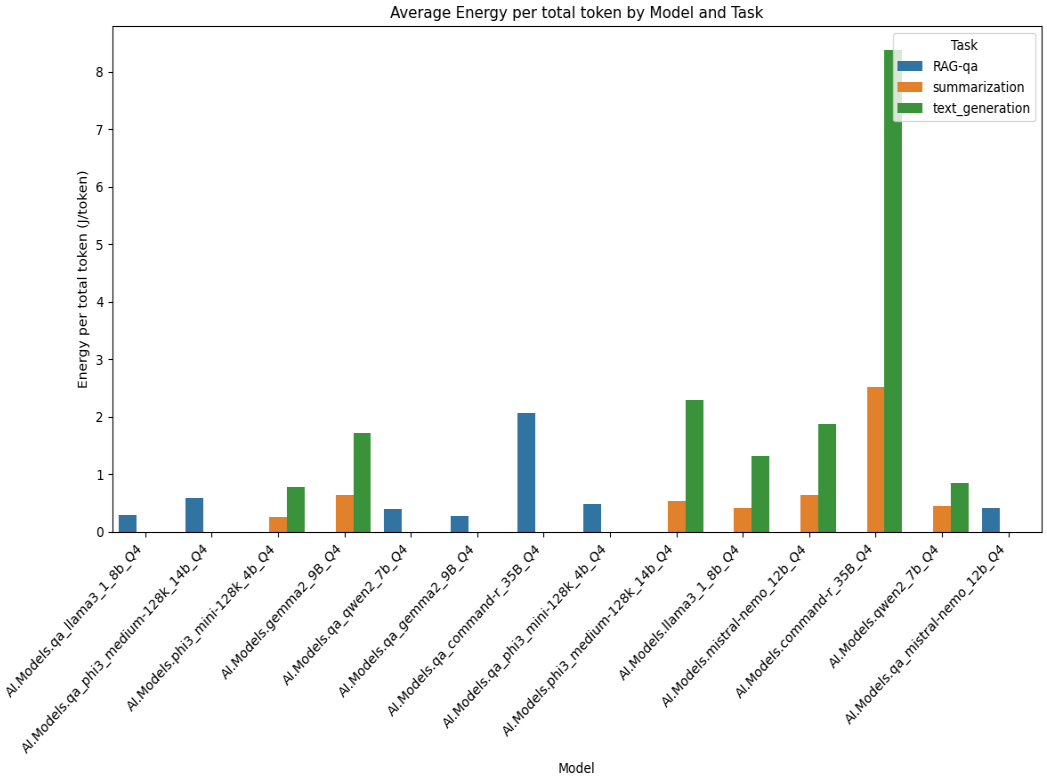

Average Power Draw by Model
Average Power Draw (W)
0
50
100
150
200
250
AI.Models.qa_llama3_1_8b_Q4
AI.Models.qa_phi3_medium-128k_14b_Q4
AI.Models.phi3_mini-128k_4b_Q4
AI.Models.gemma2_9B_Q4
AI.Models.qa_qwen2_7b_Q4
AI.Models.qa_gemma2_9B_Q4
AI.Models.qa_command-r_35B_Q4
AI.Models.qa_phi3_mini-128k_4b_Q4
AI.Models.phi3_medium-128k_14b_Q4
AI.Models.llama3_1_8b_Q4
AI.Models.mistral-nemo_12b_Q4
AI.Models.command-r_35B_Q4
AI.Models.qwen2_7b_Q4
AI.Models.qa_mistral-nemo_12b_Q4
Model


Model Average Power Consumption
Model Average Power (J)
0
5000
10000
15000
20000
25000
AI.Models.qa_llama3_1_8b_Q4
AI.Models.qa_phi3_medium-128k_14b_Q4
AI.Models.phi3_mini-128k_4b_Q4
AI.Models.gemma2_9B_Q4
AI.Models.qa_qwen2_7b_Q4
AI.Models.qa_gemma2_9B_Q4
AI.Models.qa_command-r_35B_Q4
AI.Models.qa_phi3_mini-128k_4b_Q4
AI.Models.phi3_medium-128k_14b_Q4
AI.Models.llama3_1_8b_Q4
AI.Models.mistral-nemo_12b_Q4
AI.Models.command-r_35B_Q4
AI.Models.qwen2_7b_Q4
AI.Models.qa_mistral-nemo_12b_Q4
Model


Average Energy consumption by Model and Task
Energy consumption (J)
0
5000
10000
15000
20000
25000
Task
RAG-qa
summarization
text_generation
AI.Models.qa_llama3_1_8b_Q4
AI.Models.qa_phi3_medium-128k_14b_Q4
AI.Models.phi3_mini-128k_4b_Q4
AI.Models.gemma2_9B_Q4
AI.Models.qa_qwen2_7b_Q4
AI.Models.qa_gemma2_9B_Q4
AI.Models.qa_command-r_35B_Q4
AI.Models.qa_phi3_mini-128k_4b_Q4
AI.Models.phi3_medium-128k_14b_Q4
AI.Models.llama3_1_8b_Q4
AI.Models.mistral-nemo_12b_Q4
AI.Models.command-r_35B_Q4
AI.Models.qwen2_7b_Q4
AI.Models.qa_mistral-nemo_12b_Q4
Model


Average Energy per input token by Model and Task
Energy per input token (J/token)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Task
RAG-qa
summarization
text_generation
AI.Models.qa_llama3_1_8b_Q4
AI.Models.qa_phi3_medium-128k_14b_Q4
AI.Models.phi3_mini-128k_4b_Q4
AI.Models.gemma2_9B_Q4
AI.Models.qa_qwen2_7b_Q4
AI.Models.qa_gemma2_9B_Q4
AI.Models.qa_command-r_35B_Q4
AI.Models.qa_phi3_mini-128k_4b_Q4
AI.Models.phi3_medium-128k_14b_Q4
AI.Models.llama3_1_8b_Q4
AI.Models.mistral-nemo_12b_Q4
AI.Models.command-r_35B_Q4
AI.Models.qwen2_7b_Q4
AI.Models.qa_mistral-nemo_12b_Q4
Model


Average Energy per output token by Model and Task
Energy per output token (J/token)
Model
Task
RAG-qa
summarization
text_generation
AI.Models.qa_llama3_1_8b_Q4
AI.Models.qa_phi3_medium-128k_14b_Q4
AI.Models.phi3_mini-128k_4b_Q4
AI.Models.gemma2_9B_Q4
AI.Models.qa_qwen2_7b_Q4
AI.Models.qa_gemma2_9B_Q4
AI.Models.qa_command-r_35B_Q4
AI.Models.qa_phi3_mini-128k_4b_Q4
AI.Models.phi3_medium-128k_14b_Q4
AI.Models.llama3_1_8b_Q4
AI.Models.mistral-nemo_12b_Q4
AI.Models.command-r_35B_Q4
AI.Models.qwen2_7b_Q4
AI.Models.qa_mistral-nemo_12b_Q4


Average Energy per total token by Model and Task
Energy per total token (J/token)
Model
Task
RAG-qa
summarization
text_generation
AI.Models.qa_llama3_1_8b_Q4
AI.Models.qa_phi3_medium-128k_14b_Q4
AI.Models.phi3_mini-128k_4b_Q4
AI.Models.gemma2_9B_Q4
AI.Models.qa_qwen2_7b_Q4
AI.Models.qa_gemma2_9B_Q4
AI.Models.qa_command-r_35B_Q4
AI.Models.qa_phi3_mini-128k_4b_Q4
AI.Models.phi3_medium-128k_14b_Q4
AI.Models.llama3_1_8b_Q4
AI.Models.mistral-nemo_12b_Q4
AI.Models.command-r_35B_Q4
AI.Models.qwen2_7b_Q4
AI.Models.qa_mistral-nemo_12b_Q4

Evaluation Results

Rag-qa Results

Model: AI.Models.qa_GPT4o

File: qa_001_result.json

Overall_score: 0.39796888705758016
Individual_scores:
AnswerRelevancy: 0.8333333333333334
Faithfulness: 0.875
ContextualPrecision: 0
ContextualRecall: 0.5
CustomContextualRelevancy: 0.0
Correctness: 0.17947998901214762
Reasons:
AnswerRelevancy: The score is 0.83 because the response largely addresses the issue of the missing 'Bell' icon in XWiki, but includes several irrelevant source links and a general comment about dependencies that do not directly contribute to solving the problem.
Faithfulness: The score is 0.88 because the actual output incorrectly claims that the Extension Manager can manually install the 'Notifications Application' by uploading the XAR file, whereas the retrieval context clearly states that offline installation of extensions is not supported in XWiki.
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'installing extensions using the Extension Manager' and does not mention enabling notifications or the `xwiki.properties` file, which is unrelated to the issue of the 'Bell' icon not being available. Similarly, the second node addresses 'issues related to administration access and configurable applications,' which also does not pertain to the problem at hand. None of the nodes provide relevant information on enabling notifications or fixing the bell icon issue, leading to a contextual precision score of 0.00.
ContextualRecall: The score is 0.50 because while the retrieval context mentions configuring notifications in the wiki, it lacks specific details about setting 'notifications.enabled' in 'xwiki.properties' to 'true'.
CustomContextualRelevancy: The retrieval context does not mention enabling notifications through the `xwiki.properties` file or the `notifications.enabled` setting.
Correctness: The actual output discusses troubleshooting notifications but does not specifically mention enabling them via the `notifications.enabled` setting in `xwiki.properties`, which is the key detail in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.7107189196970228
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.7
ContextualRecall: 0.5714285714285714
CustomContextualRelevancy: 0.6
Correctness: 0.3928849467535657

Reasons:
AnswerRelevancy: The score is 1.00 because the response is perfectly relevant and directly addresses the issue of not receiving notifications in XWiki. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.70 because the first node in the retrieval context is relevant, explaining 'how to follow a user to receive notifications,' and is ranked higher than irrelevant nodes. However, the second node discusses 'filters for hidden pages and other technical details,' which do not directly relate to enabling notifications, and should be ranked lower. Additionally, the fifth node, which is relevant and 'provides information about the notifications settings menu,' is ranked lower than several irrelevant nodes, such as the third node about 'customizing email templates for notifications,' which does not directly help in resolving the issue.
ContextualRecall: The score is 0.57 because while the retrieval context nodes support several aspects of the expected output, such as watching pages or users and using the alert menu, they lack information on enabling notifications in `xwiki.properties` and subscribing to pages or users, which are crucial details in the expected output.
CustomContextualRelevancy: The retrieval context covers enabling notifications via user profile and watching users, but lacks details on setting 'notifications.enabled' in 'xwiki.properties' and specific steps to watch pages or entire wikis.
Correctness: The actual output covers some aspects like watching users or pages and enabling notifications but includes many irrelevant points such as email settings, admin settings, and filtering options. It lacks direct instructions on setting 'notifications.enabled' to 'true' in 'xwiki.properties' and omits details about using the 'alert' menu.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.9414102361952148
Individual_scores:
AnswerRelevancy: 0.75
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.898461417171289
Reasons:
AnswerRelevancy: The score is 0.75 because the response addresses the main question about enabling notifications for one's own actions in XWiki, but includes multiple irrelevant source links that do not directly contribute to the solution.
Faithfulness: The score is 1.00 because there are no contradictions. Everything aligns perfectly with the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant nodes in the retrieval context are perfectly ranked higher than the irrelevant nodes. The first node states, 'By default, one don’t receive notification about one’s own activity... You have the ability to disable this filter in the “Advanced filtering options” section of your own notification settings,' directly addressing the input question. The second node also aligns with the input by mentioning, 'By default, you won’t receive notifications for actions done by yourself. This can be changed by switching off the Own event filter.' All irrelevant nodes, such as the third node discussing following a user and receiving notifications about events triggered by them, are correctly ranked lower.

ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context contains all the information from the expected output, including the default behavior of not receiving notifications about one's own activity unless explicitly targeted and the ability to disable this filter in the 'Advanced filtering options' section of the user profile.
Correctness: The actual output accurately reflects the default notification settings and the ability to change them. It mentions the default filter and steps to disable it, aligning with expected output. The only minor discrepancy is the actual output's additional detail about the 'Own event filter' naming.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.5954516738733368
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 0.8333333333333334
ContextualPrecision: 0.33730158730158727
ContextualRecall: 0.75
CustomContextualRelevancy: 0.3
Correctness: 0.3520751226051001
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question about disabling unwanted notifications in XWiki without any irrelevant information. Great job!
Faithfulness: The score is 0.83 because the actual output incorrectly claims the existence of 'Advanced filtering options' in the notifications settings, which is not supported by the retrieval context.
ContextualPrecision: The score is 0.34 because the relevant nodes in the retrieval context are ranked lower than several irrelevant nodes. For example, the first node discusses 'following a user and receiving notifications about their activities,' which is not directly related to disabling notifications for pages, yet it is ranked higher than relevant nodes. Similarly, the second node focuses on 'selecting the level of detail in email notifications' and the third node is about 'enabling or disabling the email feature,' both of which do not address the input query. Relevant nodes, such as the fourth node that mentions 'configuring notification settings,' are ranked lower, affecting the score.
ContextualRecall: The score is 0.75 because the retrieval context supports the idea of automatic page watching and changing settings in user profiles, as mentioned in the 5th node in retrieval context, but it does not specifically mention notifications about changes on your work.
CustomContextualRelevancy: The retrieval context mentions automatic watch settings and user profile notification settings but lacks details on major modifications triggering automatic page watching and deleting filters in 'Custom Filters'.
Correctness: The actual output describes how to manage notifications but lacks specific details about automatic page watching and deleting notification filters, which are present in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.9060536385671756
Individual_scores:
AnswerRelevancy: 0.9285714285714286

Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.6
CustomContextualRelevancy: 1.0
Correctness: 0.9077504028316248
Reasons:
AnswerRelevancy: The score is 0.93 because the response is highly relevant and mostly addresses the question about configuring OpenID Connect in XWiki. However, there is a minor irrelevant mention of LDAP Authenticator, which slightly detracts from the focus on OpenID Connect configuration.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant node, which mentions '# oidc.enableUser=true' as the necessary configuration to enable user accounts on first login, is ranked first. All irrelevant nodes, which discuss topics like domain-based authentication, session cookies, password resetting, and other unrelated authentication mechanisms, are ranked lower, ensuring precise and accurate retrieval.
ContextualRecall: The score is 0.60 because while the retrieval context's 1st node supports the instructions regarding the 'oidc.enableUser=true' property in the xwiki.properties file, it lacks explicit details about saving the file, restarting the XWiki instance, and ensuring new user accounts are automatically enabled on first login via OpenID Connect.
CustomContextualRelevancy: The retrieval context contains the exact information from the expected answer, specifically mentioning the property 'oidc.enableUser=true' in the 'xwiki.properties' file, which matches the expected output.
Correctness: The actual output accurately reflects the expected instructions for setting `oidc.enableUser=true` and restarting XWiki. However, it does not mention uncommenting the property, which is noted in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.9145316594031131
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.6
CustomContextualRelevancy: 1.0
Correctness: 0.8871899564186785
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the first node in the retrieval context is directly related to customizing user profile fields, mentioning 'It's possible to associate non standard properties coming from the OpenID Connect provider with the XWiki user' and providing an example of mapping. All irrelevant nodes, such as the second node discussing 'access tokens and scopes' and the third node focused on 'group synchronization and custom claims', are ranked lower, ensuring high precision. Great job maintaining perfect precision!

ContextualRecall: The score is 0.60 because while the main concept of mapping user profile fields from the OpenID Connect provider is supported by nodes in the retrieval context, specific instructions such as replacing placeholders and restarting XWiki are not explicitly found in the context.
CustomContextualRelevancy: The retrieval context contains all the information from the expected answer, including customizing user profile fields using `oidc.user.mapping` in `xwiki.properties` and the example mappings provided.
Correctness: The actual output correctly describes the customization of user profile fields through the 'oidc.user.mapping' property and provides examples of mapping, following the expected output closely. However, it includes an additional note on using transformations like '.lowerCase', which is not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.8071370102121169
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.8333333333333334
ContextualRecall: 0.5
CustomContextualRelevancy: 0.8
Correctness: 0.7094887279393676
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question about configuring group synchronization between XWiki and an OpenID Connect provider without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.83 because relevant nodes discussing the `oidc.groups.mapping` and `oidc.groups.claim` properties, crucial for configuring group synchronization, are ranked higher. However, the score is not higher because nodes in retrieval context ranked third, fourth, and fifth provide general information about OpenID Connect, authentication management, and dependencies, which are not directly relevant to the group synchronization process.
ContextualRecall: The score is 0.50 because while there are supportive nodes in the retrieval context for configuring properties like `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping`, key procedural steps such as locating and editing the `xwiki.properties` file, and saving and restarting XWiki are not supported by any nodes in the retrieval context.
CustomContextualRelevancy: The retrieval context includes details on configuring `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping`, which align with the expected output. However, it lacks explicit instructions on locating and editing the `xwiki.properties` file and restarting the XWiki instance.
Correctness: The actual output includes configuring `oidc.groups.claim`, `oidc.groups.mapping`, and `oidc.groups.separator`, aligning with expected output. However, it adds extra steps like enabling group synchronization and handling group restrictions, which are not mentioned in the expected output. The step about saving and restarting XWiki is missing.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.6091088133632122
Individual_scores:
AnswerRelevancy: 0.9
Faithfulness: 1.0
ContextualPrecision: 0.2
ContextualRecall: 0.5
CustomContextualRelevancy: 0.3
Correctness: 0.754652880179273
Reasons:
AnswerRelevancy: The score is 0.90 because the output mostly addresses the question about setting upload restrictions, but includes links to general XWiki concepts and configurations that aren't directly related to the specific requirement of restricting attachments to images or PDFs of a certain size.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.20 because the relevant node, which 'mentions setting limits on the maximum size of an attachment and provides steps to change it', is ranked fifth. This indicates that irrelevant nodes, such as the first node discussing 'linking to external files and images', the second node providing 'a general overview of XWiki's structure and features', and the third node explaining 'how to add attachments using the XWiki interface', are ranked higher than the relevant information. These irrelevant nodes do not address configuring upload restrictions or size limits, which is why the score is not higher.
ContextualRecall: The score is 0.50 because while some nodes in the retrieval context provide relevant information about configuring xwiki.properties and setting maximum attachment size (nodes 5 and 6), they do not fully cover all the steps and details mentioned in the expected output, such as the specific line for mimetype restrictions and saving changes for subwikis.
CustomContextualRelevancy: The retrieval context mentions setting maximum attachment size and mimetype restrictions but lacks specific details on configuring xwiki.properties or setting size to 10MB as in the expected output.
Correctness: The actual output correctly details setting the maximum upload size and mentions mimetype restrictions but suggests using the Attachment Validation Application instead of directly editing the xwiki.properties file for mimetype restrictions, potentially misleading from the expected method.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.6400406063916423
Individual_scores:
AnswerRelevancy: 0.8095238095238095
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.030719828826044033
Reasons:
AnswerRelevancy: The score is 0.81 because the response mostly addresses the question about denying script rights to a space administrator, but it includes multiple irrelevant source links that do not contribute to answering the question directly.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!

ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'programming rights and their implications', but does not address the denial of script rights to a space administrator. The second node lists 'various rights and their default states', but does not mention the inability to deny script rights to a space administrator. The third node provides 'descriptions of rights', but does not address the specific issue of denying script rights to a space administrator. Subsequent nodes similarly fail to address the input query, focusing on unrelated topics such as 'comment, edit, and delete rights', 'technical implementation of access checks', and 'converting rights during page migration'.
ContextualRecall: The score is 1.00 because all sentences in the expected output are fully supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context confirms that rights implied by admin cannot be denied, matching the expected output.
Correctness: The actual output incorrectly suggests that script rights can be denied for a space administrator, contradicting the expected output that states this is not supported in XWiki.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.6409501522286888
Individual_scores:
AnswerRelevancy: 0.7142857142857143
Faithfulness: 1.0
ContextualPrecision: 0.16666666666666666
ContextualRecall: 1.0
CustomContextualRelevancy: 0.2
Correctness: 0.7647485324197508
Reasons:
AnswerRelevancy: The score is 0.71 because the output partially addresses the input question about configuring permissions in XWiki, but includes multiple irrelevant source links that do not contribute to answering the question.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.17 because the relevant node in retrieval context, which provides information on XWiki's permission system including setting view and edit rights, is ranked sixth. The first node discusses content organization and rights inheritance in XWiki, which doesn't directly address setting view and edit permissions for teams. The second node is about the parent/child relationship and nested spaces, not directly related to configuring team-specific permissions. The third node discusses string formatting and user properties in OpenID Connect, unrelated to access control in XWiki. These irrelevant nodes should be ranked lower than the relevant node.
ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context mentions setting rights at different levels and inheritance, but lacks specifics on creating groups for each team, setting wiki-wide rights, and space-level rights as described in the expected output.
Correctness: The actual output accurately describes setting global view permissions and restricting edit permissions to specific spaces, aligning with the expected output. However, it lacks mention of creating groups for each team and inheriting rights for child pages, which are key facts in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.7537122941725168
Individual_scores:
AnswerRelevancy: 0.6875
Faithfulness: 0.8333333333333334
ContextualPrecision: 0.5
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.5014404317017673
Reasons:
AnswerRelevancy: The score is 0.69 because the output includes relevant information about authenticating users with access tokens, but it is lowered due to repeated irrelevant statements providing source links that do not directly address the question.
Faithfulness: The score is 0.83 because the actual output incorrectly suggests setting `xwiki.authentication.authclass` to `org.xwiki.contrib.llm.authentication.JWTTokenAuth` for JWT Token Authentication, whereas the retrieval context specifies it should be set to `com.xpn.xwiki.user.impl.xwiki.AppServerTrustedAuthServiceImpl` for container authentication.
ContextualPrecision: The score is 0.50 because relevant nodes in the retrieval context are not consistently ranked higher than irrelevant nodes. For example, the first node is about 'LDAP configuration' and does not mention token-based authentication, which should be ranked lower. Similarly, the third node is about 'LDAP authentication' and is not related to token-based authentication, yet it appears before more relevant nodes. On the other hand, relevant nodes like the second node, which discusses 'configuration of authorized applications' for token-based authentication, and the fourth node, which details 'JWT token generation', are appropriately included but not consistently prioritized over irrelevant nodes.
ContextualRecall: The score is 1.00 because every sentence in the expected output is fully supported by the relevant nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context contains detailed information about token-based authentication, JWT claims, and configuration steps, matching the expected answer.
Correctness: The actual output mentions JWT Token Authentication and setting xwiki.authentication.authclass, aligning with the expected output. However, it introduces OpenID Connect which is not in the expected output, and lacks details on configuring authorized applications and claim requirements.
Question_language: en
Expected_answer_language: en
Answer_language: en

Model: AI.Models.qa_GPT4o-mini

File: qa_001_result.json

Overall_score: 0.45158911416588193
Individual_scores:
AnswerRelevancy: 0.8518518518518519
Faithfulness: 1.0
ContextualPrecision: 0.30952380952380953
ContextualRecall: 0.5
CustomContextualRelevancy: 0.0
Correctness: 0.04815902361963014
Reasons:

AnswerRelevancy: The score is 0.85 because the response is mostly relevant to the issue of the missing 'Bell' icon in XWiki, but it includes repeated irrelevant statements about source links that do not directly address the problem.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.31 because the relevant nodes in the retrieval context are ranked lower than several irrelevant nodes. For instance, the first node discusses 'installation instructions and prerequisites', which is not relevant to fixing the bell issue, and the second node is about 'access and error messages related to configurable applications', which also does not pertain to the problem. The third node, which is relevant as it mentions the 'Notifications Application' and the bell icon, is ranked after these irrelevant nodes. Similarly, the seventh node, which is relevant because it discusses 'configuring notification settings in the administration', is also ranked lower than irrelevant nodes like the fourth node, which only describes 'the content of the Alert menu and notifications' without addressing the bell issue.
ContextualRecall: The score is 0.50 because while the 7th node in the retrieval context mentions 'notifications' and 'xwiki.properties', there is no specific mention of enabling notifications via 'xwiki.properties', which is crucial for fully supporting the expected output.
CustomContextualRelevancy: The retrieval context does not mention enabling notifications via the `xwiki.properties` file, which is the key information in the expected output.
Correctness: The actual output does not mention enabling notifications via `notifications.enabled` in `xwiki.properties`, and introduces unrelated troubleshooting steps.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.6841433173428224
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.6666666666666666
ContextualRecall: 0.5
CustomContextualRelevancy: 0.6
Correctness: 0.3381932373902682
Reasons:
AnswerRelevancy: The score is 1.00 because the response is perfectly relevant and addresses the question without any irrelevant statements. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.67 because the first node in the retrieval context is relevant, explaining how to follow a user and receive notifications, which is crucial for ensuring notifications are set up correctly. However, the second node discusses 'filters and hidden pages,' which is not directly related to enabling notifications, and should be ranked lower. Additionally, the third node focuses on 'customizing the notification email template,' which is not directly relevant to the issue of not receiving notifications, and should also be ranked lower than relevant nodes.
ContextualRecall: The score is 0.50 because while the retrieval context supports details about watching pages, following users, and using the 'network' tab (as seen in node 1 of the retrieval context), it lacks information about enabling notifications in 'xwiki.properties' and subscribing to pages or users for notifications.
CustomContextualRelevancy: The retrieval context covers enabling notifications and following users, but lacks details on setting 'notifications.enabled' in 'xwiki.properties' and subscribing to pages.

Correctness: The actual output mentions enabling notifications and watching users or pages, but lacks specific instructions on editing `xwiki.properties` and does not address using the 'alert' menu or checking the 'network' tab.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.9496539698274892
Individual_scores:
AnswerRelevancy: 0.7692307692307693
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.9286930497341664
Reasons:
AnswerRelevancy: The score is 0.77 because the response mostly addresses the question about enabling notifications for one's own actions in XWiki. However, the inclusion of source links that do not directly answer the question slightly detracts from the overall relevancy.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than the irrelevant ones. The first two nodes directly address the input by explaining how to enable notifications for one's own actions, while the irrelevant nodes discuss unrelated topics such as following a user, administrator settings, RSS feeds, and notification methods, which are not pertinent to the input question. Great job on the perfect ranking!
ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context fully contains the expected answer, including the default behavior of not receiving notifications about one's own activity unless explicitly targeted, and the ability to disable this filter in the 'Advanced filtering options' section of the user profile.
Correctness: The actual output accurately reflects the expected output by explaining that users do not receive notifications for their own actions by default and provides instructions on how to disable the filter in the 'Advanced filtering options' of notification settings, matching the expected output details.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.5707448917994089
Individual_scores:
AnswerRelevancy: 0.8947368421052632
Faithfulness: 1.0
ContextualPrecision: 0.26785714285714285
ContextualRecall: 0.75
CustomContextualRelevancy: 0.3
Correctness: 0.21187536583404754
Reasons:

AnswerRelevancy: The score is 0.89 because the response mostly addresses how to disable notifications in XWiki, but includes a source link and a standalone 'Filters:' statement that do not directly contribute to answering the question.
Faithfulness: The score is 1.00 because there are no contradictions. Great job on maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 0.27 because the relevant nodes in the retrieval context are ranked lower than several irrelevant nodes. For instance, the first node discusses 'following a user and default watch settings,' which does not address disabling notifications, while the fourth node, which is relevant, mentions 'administrators can change users' notification settings.' Additionally, the second node focuses on 'details of changes and macros for settings,' which is not pertinent, whereas the seventh node, which is relevant, discusses 'filtering notifications, including advanced filtering options.' These relevant nodes should be ranked higher to improve the score.
ContextualRecall: The score is 0.75 because most of the expected output is supported by node 5 in the retrieval context, which covers automatic page watching, changing settings, and custom filters. However, the retrieval context does not mention notifications about changes to your work specifically, leading to a partial match.
CustomContextualRelevancy: The retrieval context mentions default watch settings and notification settings in user profiles, but lacks details about automatic page watching for major modifications and deleting notification filters.
Correctness: The actual output provides steps for managing notifications but lacks key facts about automatic page watching and deleting custom filters from the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.8990661087973559
Individual_scores:
AnswerRelevancy: 0.9166666666666666
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.6
CustomContextualRelevancy: 1.0
Correctness: 0.8777299861174692
Reasons:
AnswerRelevancy: The score is 0.92 because the response is highly relevant to configuring OpenID Connect in XWiki, but it includes an irrelevant mention of LDAP authentication, which slightly detracts from the overall focus.
Faithfulness: The score is 1.00 because there are no contradictions, indicating a perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which discusses the property '# oidc.enableUser=true' necessary for enabling user accounts on first login, is ranked first. This ensures that the most pertinent information is prioritized over irrelevant nodes discussing topics like domain-based authentication, password recovery, and general OpenID Connect configurations.
ContextualRecall: The score is 0.60 because while the 1st node in the retrieval context supports the instructions to adjust the `xwiki.properties` file and set the `oidc.enableUser` property to `true`, it lacks information about saving the file, restarting the XWiki instance, and ensuring new user accounts are automatically enabled on first login via OpenID Connect.
CustomContextualRelevancy: The retrieval context contains the exact information from the expected output, specifically mentioning the property `oidc.enableUser=true` in the `xwiki.properties` file, which

matches the expected answer.
Correctness: The actual output correctly describes adjusting the `oidc.enableUser` property, saving the file, and restarting XWiki. It omits the specific commenting/uncommenting detail present in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.8490539647125716
Individual_scores:
AnswerRelevancy: 0.8823529411764706
Faithfulness: 1.0
ContextualPrecision: 0.625
ContextualRecall: 0.8
CustomContextualRelevancy: 1.0
Correctness: 0.7869708470989581
Reasons:
AnswerRelevancy: The score is 0.88 because the response is mostly relevant to the question about customizing user profile fields in XWiki using OpenID Connect. However, it includes citations and suggestions to consult documentation, which do not directly address the specific customization query, preventing a higher score.
Faithfulness: The score is 1.00 because there are no contradictions. Great job maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 0.62 because the first node in the retrieval context is relevant and directly addresses customizing user profile fields with an example of mapping, which is highly pertinent to the input. However, several irrelevant nodes, such as the second node discussing 'access tokens and scopes', the third node about 'group claims and synchronization', and others, are ranked higher than the last relevant node, which lists available variables for mapping. These irrelevant nodes should be ranked lower to improve the precision score.
ContextualRecall: The score is 0.80 because the retrieval context nodes support most of the expected output, particularly the customization and mapping of user profile fields in the `xwiki.properties` file (sentence 1-4). However, it lacks information on saving the file and restarting XWiki to apply changes (sentence 5).
CustomContextualRelevancy: The retrieval context contains the exact information from the expected output, including the ability to map user profile fields using the `oidc.user.mapping` property in the `xwiki.properties` file with examples provided.
Correctness: The actual output accurately describes customizing user profile fields using `oidc.user.mapping` and provides example mappings. However, it includes additional information about using custom claims and consulting documentation, which is not present in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.7763624843120526
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.8333333333333334

ContextualRecall: 0.5
CustomContextualRelevancy: 0.8
Correctness: 0.5248415725389818
Reasons:
AnswerRelevancy: The score is 1.00 because the output is perfectly relevant and directly addresses the question about configuring group synchronization between XWiki and an OpenID Connect provider. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.83 because the relevant nodes in the retrieval context are mostly ranked higher than the irrelevant ones. The first and second nodes are relevant as they discuss 'oidc.groups.mapping' and group synchronization, which are crucial for configuring group synchronization between XWiki and an OpenID Connect provider. However, the third node, which provides general information about OpenID Connect without addressing group synchronization, should be ranked lower. Similarly, the fourth node focuses on authentication management and endpoints, and the fifth node lists dependencies, both of which are not directly related to the group synchronization process. The sixth node is relevant as it describes the 'oidc.groups.claim' property, essential for the setup, and is appropriately ranked.
ContextualRecall: The score is 0.50 because while several specific configuration steps (sentences 3, 5, and 7) are supported by nodes in the retrieval context, many general instructions and replacement examples (sentences 1, 2, 4, 6, 8, and 9) are not directly supported by any nodes in the retrieval context.
CustomContextualRelevancy: The retrieval context includes details on configuring 'oidc.groups.claim', 'oidc.groups.separator', and 'oidc.groups.mapping', which align with the expected output. However, it lacks explicit instructions on locating and editing the 'xwiki.properties' file and restarting the XWiki instance.
Correctness: The actual output mentions configuring `oidc.groups.claim` and `oidc.groups.mapping`, aligning with expected output. However, it lacks details on `oidc.groups.separator` and saving/restarting XWiki. Additional steps like enabling group synchronization and testing are present but not in expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.6273544245715755
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.2
ContextualRecall: 0.5833333333333334
CustomContextualRelevancy: 0.3
Correctness: 0.6807932140961195
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions. Great job on maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 0.20 because the relevant node in the retrieval context, which 'mentions setting limits on the maximum size of an attachment and provides steps to change it', is ranked fifth. However, several irrelevant nodes are ranked higher, such as the first node, which

'discusses linking to external files and images, which is unrelated to configuring upload restrictions for attachments', and the second node, which 'provides general information about XWiki pages and their components, not about configuring attachment upload settings'.
ContextualRecall: The score is 0.58 because while several steps related to configuring maximum attachment size (Step 2) are well-supported by node 5 in the retrieval context, key instructions regarding mimetype configuration (Step 1) and ensuring only specific file types and sizes are allowed lack direct support from any node in the retrieval context.
CustomContextualRelevancy: The retrieval context mentions setting maximum attachment size and configuring mimetype restrictions, but lacks specifics on editing xwiki.properties for mimetypes and setting 10MB limit.
Correctness: The actual output correctly identifies the need to set a 10MB upload limit and specify allowed MIME types but suggests using the Attachment Validation Application instead of directly editing `xwiki.properties` for MIME types. It also includes additional steps like security settings and testing, which are not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.7189866824745721
Individual_scores:
AnswerRelevancy: 0.8
Faithfulness: 1.0
ContextualPrecision: 0.7
ContextualRecall: 1.0
CustomContextualRelevancy: 0.8
Correctness: 0.01392009484743239
Reasons:
AnswerRelevancy: The score is 0.80 because the answer mostly addresses the question about denying script rights to a space administrator, but it includes multiple irrelevant source links that do not directly pertain to the question.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.70 because the first node in the retrieval context is relevant, as it mentions that 'programming rights can only be granted from the wiki preferences page,' implying that certain rights, including admin rights, cannot be denied once granted. However, the second node is irrelevant, as it primarily lists various rights and their default states but does not specifically address the denial of script rights to a space administrator. The third node is also irrelevant, providing a tabular view of rights but not mentioning the inability to deny script rights to a space administrator. These irrelevant nodes are ranked higher than the fifth node, which is relevant and explains that 'the script right gives a lot of power to users' and that administrators have to manually allow it, suggesting that once allowed, it cannot be denied.
ContextualRecall: The score is 1.00 because the expected output is fully supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context mentions that admin rights imply other rights and are not deniable, aligning with the expected output that rights implied by admin cannot be denied. However, it does not explicitly state that denying admin rights is unsupported in XWiki.
Correctness: The actual output provides steps to deny script rights to a space administrator in XWiki, contradicting the expected output which states it is not supported and rights implied by admin right cannot be denied.
Question_language: en

Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.6467811474647401
Individual_scores:
AnswerRelevancy: 0.8
Faithfulness: 0.8888888888888888
ContextualPrecision: 0.20833333333333331
ContextualRecall: 0.875
CustomContextualRelevancy: 0.3
Correctness: 0.8084646625662186
Reasons:
AnswerRelevancy: The score is 0.80 because the response effectively addresses the main question about configuring access rights in XWiki, but the inclusion of multiple source links, which are not directly relevant to the configuration process, slightly detracts from the overall relevancy.
Faithfulness: The score is 0.89 because the actual output incorrectly states that space-level permissions override global permissions, whereas the retrieval context clarifies that page-level permissions can override wiki-wide permissions.
ContextualPrecision: The score is 0.21 because the relevant nodes in the retrieval context are ranked lower than several irrelevant nodes. The first node discusses 'the organization of content in XWiki and the concept of nested pages,' which does not provide guidance on setting view and edit rights for teams. The second node explains 'the parent/child relationship and nested spaces,' but does not address configuring access rights for teams. The third node is about 'string formatting and user identifiers,' which is unrelated to the input question. The sixth node, which is relevant, provides information on XWiki's permission system, including setting view and edit rights. The eighth node, also relevant, outlines basic rules for setting rights in XWiki, including creating groups and setting page-level rights. These relevant nodes should be ranked higher to improve the score.
ContextualRecall: The score is 0.88 because most of the expected output sentences are well supported by the nodes in the retrieval context, particularly node 6, which covers key aspects like creating groups, setting wiki-wide and space-level rights, and ensuring view and edit permissions. However, the retrieval context lacks information on adjusting rights for additional teams or spaces, which slightly impacts the score.
CustomContextualRelevancy: The retrieval context mentions rights inheritance and setting rights at different levels, but lacks specific steps like creating groups for each team or setting 'Edit' rights for team spaces.
Correctness: The actual output accurately describes setting global view rights and specific edit rights, ensuring teams can view all spaces but only edit their own. However, it lacks specific mention of creating groups for each team and inheriting rights for child pages as outlined in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.8156800938781242
Individual_scores:
AnswerRelevancy: 0.7058823529411765
Faithfulness: 1.0
ContextualPrecision: 0.5
ContextualRecall: 1.0

CustomContextualRelevancy: 0.9
Correctness: 0.7881982103275684
Reasons:
AnswerRelevancy: The score is 0.71 because the answer partially addresses the question about authenticating users with access tokens, but includes multiple irrelevant statements providing source links that do not directly answer the question.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.50 because relevant nodes in the retrieval context are not consistently ranked higher than irrelevant ones. For instance, the first node discusses 'LDAP authentication settings, which are unrelated to token-based authentication using JWTs,' and should be ranked lower. Similarly, the third node 'describes a method for checking authentication but does not relate to token-based authentication or access tokens,' yet it appears before relevant nodes. However, relevant nodes like the second node, which 'provides information about configuring authorized applications, which is relevant for using access tokens for authentication,' are correctly identified, contributing to the current score.
ContextualRecall: The score is 1.00 because every sentence in the expected output is fully supported by the nodes in the retrieval context, demonstrating a perfect match. Great job!
CustomContextualRelevancy: The retrieval context includes information about enabling the token authenticator, configuring authorized applications, and generating JWT tokens with required claims and Ed25519 key, aligning closely with the expected output. However, it lacks explicit mention of the limitation regarding tokens for existing users not created through this authenticator.
Correctness: The actual output correctly describes the authentication process using JWT for the LLM Application and matches most facts from the expected output, including the configuration and token requirements. However, it includes additional information on fallback authentication not present in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

Model: AI.Models.qa_claude3_5_sonet

File: qa_001_result.json

Overall_score: 0.3478083348371878
Individual_scores:
AnswerRelevancy: 0.9259259259259259
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 0.0
CustomContextualRelevancy: 0.0
Correctness: 0.1609240830972009
Reasons:
AnswerRelevancy: The score is 0.93 because the response is highly relevant and provides useful information on addressing the issue with the bell icon in XWiki. However, it includes a source link that does not directly contribute to solving the problem, which slightly impacts the score.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'installation instructions and prerequisites', which do not relate to enabling notifications in the wiki. The second node describes 'the notifications application and its features', but does not address enabling notifications. The third node lists 'dependencies for the

notifications extension', lacking information on enabling notifications. The fourth node talks about 'overriding notification templates', which is unrelated to enabling notifications in the wiki.
ContextualRecall: The score is 0.00 because the retrieval context does not contain any information about enabling notifications in the wiki or the `notifications.enabled` setting in `xwiki.properties`, which are crucial for the expected output.
CustomContextualRelevancy: The retrieval context does not mention enabling notifications via the `notifications.enabled` setting in `xwiki.properties`.
Correctness: The actual output provides additional installation details irrelevant to enabling notifications via `xwiki.properties`. It omits the specific instruction to set `notifications.enabled` to `true`.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.6883072105344747
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.75
ContextualRecall: 0.5714285714285714
CustomContextualRelevancy: 0.5
Correctness: 0.30841469177827696
Reasons:
AnswerRelevancy: The score is 1.00 because the response is perfectly relevant and directly addresses the issue of not receiving notifications in XWiki. Great job on providing a focused and helpful answer!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.75 because the first and fourth nodes in the retrieval context are relevant, discussing how to follow a user and various methods of receiving notifications in XWiki, which are crucial for setting up notifications correctly. However, the second node, ranked higher than the fourth, only discusses filters for notifications without addressing enabling notifications or subscriptions. Additionally, the third node, which talks about RSS feeds and notification settings, does not specifically address enabling notifications or subscribing to content, and the fifth node, discussing the 'What's New' application, is not directly related to the issue of enabling notifications.
ContextualRecall: The score is 0.57 because while the nodes in retrieval context cover details about enabling notifications through the alert menu, following users, and checking the network tab (sentences 3, 4, 5, and 6), they do not mention enabling notifications in `xwiki.properties`, subscribing to pages or users, or watching a page or its children (sentences 1, 2, and 3).
CustomContextualRelevancy: The retrieval context covers following users and some notification settings but lacks details on enabling notifications via `xwiki.properties` and watching pages or wikis.
Correctness: The actual output provides guidance on enabling notifications but lacks specific instructions on setting `notifications.enabled` in `xwiki.properties` to `true` and does not mention following users through the User Index. It includes additional information about version-specific settings and advanced filtering, which are not in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.9571824315495626

Individual_scores:
AnswerRelevancy: 0.8
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.9430945892973759
Reasons:
AnswerRelevancy: The score is 0.80 because the output mostly addresses how to enable notifications for one's own actions in XWiki, but it includes irrelevant source and image links that do not directly contribute to answering the question.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant nodes in the retrieval context are perfectly ranked higher than the irrelevant nodes. The first and second nodes directly address how to enable notifications for one's own actions, mentioning the ability to disable the filter in the 'Advanced filtering options' and switching off the 'Own event filter'. The irrelevant nodes, such as the third node discussing administrative settings and user preferences without addressing the specific question, are correctly ranked lower. Great job on achieving perfect precision!
ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context fully contains the expected answer, including the default behavior of not receiving notifications about one's own activity unless explicitly targeted and the ability to change this in the 'Advanced filtering options' of the notification settings in the user profile.
Correctness: The actual output accurately reflects the expected output by stating that users don't receive notifications about their own activity by default and that the setting can be changed in the 'Advanced filtering options' section of the notification settings. No factual discrepancies or errors are introduced.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.6904237839951525
Individual_scores:
AnswerRelevancy: 0.9166666666666666
Faithfulness: 1.0
ContextualPrecision: 0.7
ContextualRecall: 1.0
CustomContextualRelevancy: 0.3
Correctness: 0.22587603730424927
Reasons:
AnswerRelevancy: The score is 0.92 because the response is mostly relevant and provides useful information on disabling notifications in XWiki. However, it includes a source link and a heading that do not directly address the question, slightly reducing the score.
Faithfulness: The score is 1.00 because there are no contradictions; the actual output perfectly aligns with the retrieval context. Great job!
ContextualPrecision: The score is 0.70 because the first node in the retrieval context is relevant, discussing 'settings' in the notifications menu, which aligns with changing notification settings in the user profile. However, irrelevant nodes, such as the second node discussing 'following a user and default watch settings,' and the third node about 'filtering notifications,' are ranked higher than another

relevant node, the fifth node, which mentions 'default auto-watched mode' and user preferences.
ContextualRecall: The score is 1.00 because every sentence in the expected output is fully supported by the nodes in the retrieval context. Great job on aligning the information perfectly!
CustomContextualRelevancy: The retrieval context mentions default notification settings and user profile settings but lacks specific details about automatic page watching and deleting filters in the 'Custom Filters' list.
Correctness: Actual output does not mention automatic page watching or deleting filters from 'Custom Filters'. It discusses default behaviors in different versions and managing notifications, which diverges from expected focus on automatic watching and filter deletion.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.8450151287079777
Individual_scores:
AnswerRelevancy: 0.9
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.4
CustomContextualRelevancy: 1.0
Correctness: 0.7700907722478658
Reasons:
AnswerRelevancy: The score is 0.90 because the response is largely relevant and helpful in addressing the configuration of OpenID Connect in XWiki. However, it includes a mention of LDAP Authenticator, which is not pertinent to the question about OpenID Connect, preventing a perfect score.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant node, which mentions the property '# oidc.enableUser=true' directly related to enabling user accounts on first login, is ranked first. This shows a perfect ranking where relevant information is prioritized over irrelevant contexts discussing domain-based authentication, general OpenID Connect information, various authentication methods, and LDAP settings.
ContextualRecall: The score is 0.40 because while the retrieval context's 1st node contains relevant information about enabling user accounts with the property '# oidc.enableUser=true', it lacks details on adding or updating properties, saving the file, restarting XWiki, and the automatic enabling of accounts via OpenID Connect.
CustomContextualRelevancy: The retrieval context contains the exact information found in the expected answer, specifically the mention of the 'oidc.enableUser=true' property in the 'xwiki.properties' file.
Correctness: Actual output correctly identifies the need to set oidc.enableUser=true in xwiki.properties, but lacks specifics about uncommenting the line and restarting XWiki. Provides additional unnecessary information about file location and default settings.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.9259716161731638
Individual_scores:

AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.8
CustomContextualRelevancy: 1.0
Correctness: 0.7558296970389831
Reasons:
AnswerRelevancy: The score is 1.00 because the output is perfectly relevant and directly addresses the question about customizing user profile fields in XWiki using OpenID Connect. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which discusses 'oidc.user.mapping=myxproperty1=${oidc.user.subject}' directly related to mapping additional user profile fields from the OpenID Connect provider to XWiki user properties, is ranked first. This ensures that the most pertinent information is prioritized, resulting in a perfect score. Great job!
ContextualRecall: The score is 0.80 because most of the expected output is well-supported by the 1st node in the retrieval context, particularly the customization and mapping instructions. However, the instruction to restart XWiki after saving changes is not supported by the retrieval context.
CustomContextualRelevancy: The retrieval context contains all the information from the expected output, including the ability to map custom user profile fields using the 'oidc.user.mapping' property in the 'xwiki.properties' file.
Correctness: The actual output correctly describes how to use `oidc.user.mapping` in `xwiki.properties`, similar to the expected output. It includes examples that match the expected format. However, it introduces additional details about suffixes and custom listeners, which are not mentioned in the expected output but do not contradict it.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.811418661148552
Individual_scores:
AnswerRelevancy: 0.9583333333333334
Faithfulness: 0.9230769230769231
ContextualPrecision: 1.0
ContextualRecall: 0.5
CustomContextualRelevancy: 0.9
Correctness: 0.5871017104810554
Reasons:
AnswerRelevancy: The score is 0.96 because the response is highly relevant and provides a detailed explanation of configuring group synchronization between XWiki and an OpenID Connect provider. The only minor issue is the inclusion of a source link, which slightly detracts from the directness of the answer.
Faithfulness: The score is 0.92 because the actual output incorrectly claims that the default claim name for groups is `xwiki_groups`, while the retrieval context only mentions that group synchronization is enabled by adding `xwiki_groups` to the `oidc.userinfoclaims` property, without specifying it as a default claim name.
ContextualPrecision: The score is 1.00 because all relevant nodes are perfectly ranked higher than irrelevant nodes. The first three nodes in the retrieval context provide detailed information on configuring the `oidc.groups.claim` and `oidc.groups.mapping` properties, which are crucial for setting up group synchronization as required. The irrelevant nodes, ranked fourth and fifth, discuss

prerequisites, installation instructions, and authentication management, which are not directly related to the specific steps for group synchronization.
ContextualRecall: The score is 0.50 because while nodes in the retrieval context provide detailed information on configuring properties like `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping`, they lack instructions on locating, opening, and saving the `xwiki.properties` file, as well as replacing placeholders with actual values.
CustomContextualRelevancy: The retrieval context contains detailed information on configuring group synchronization, including setting the oidc.groups.claim, oidc.groups.separator, and oidc.groups.mapping properties, which matches the expected output. However, instructions to locate and edit the xwiki.properties file and restart the XWiki instance are not explicitly mentioned.
Correctness: The actual output includes details on enabling group synchronization and provides examples for configuring group claims and mapping, aligning partially with expected output. However, it omits step-by-step instructions on locating and editing `xwiki.properties`, and lacks information on saving changes and restarting XWiki.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.6226764970842024
Individual_scores:
AnswerRelevancy: 0.9411764705882353
Faithfulness: 1.0
ContextualPrecision: 0.3333333333333333
ContextualRecall: 0.6
CustomContextualRelevancy: 0.3
Correctness: 0.5615491785836456
Reasons:
AnswerRelevancy: The score is 0.94 because the response is highly relevant and mostly addresses the input question effectively. However, it could not be higher due to the vague statement 'Additional recommendations:' which does not directly contribute to the solution.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.33 because the relevant node in the retrieval context, ranked third, explains how to set the maximum size of an attachment and mentions the XWikiPreferences document, which is relevant to configuring upload size limits. However, irrelevant nodes ranked first and second discuss linking to external files and images, and provide an overview of XWiki's structure and features, which do not address attachment upload restrictions or size limits. These should be ranked lower than the relevant node.
ContextualRecall: The score is 0.60 because while several steps related to configuring attachment properties and maximum upload size in XWiki are well-supported by specific nodes in the retrieval context, some introductory and explanatory sentences do not directly match any content in the retrieval context.
CustomContextualRelevancy: The retrieval context mentions setting maximum attachment size and mimetype restrictions but lacks specific details on configuring 'xwiki.properties' for mimetypes and the exact steps for setting the size to 10MB.
Correctness: The actual output correctly sets the size limit to 10MB and provides steps to configure it, aligning with the expected output. However, it deviates on file type restriction by suggesting the use of an external application instead of configuring the xwiki.properties file directly, missing the specific configuration step.
Question_language: en

Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.6577856181566387
Individual_scores:
AnswerRelevancy: 0.8125
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.13421370893983203
Reasons:
AnswerRelevancy: The score is 0.81 because the main content addresses the question about denying script rights to a space administrator, but it includes irrelevant source links that do not directly address the question, preventing a higher score.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'programming rights and script execution' but does not address denying script rights to a space administrator. The second node provides a 'table of rights and their default states' without mentioning the inability to deny rights implied by admin rights. The third node explains the 'script right and its availability' but fails to mention denying rights. The fourth node contains 'code snippets related to checking access rights' without addressing the specific issue. Lastly, the fifth node discusses 'converting rights and excluding pages' which is unrelated to the input query.
ContextualRecall: The score is 1.00 because the expected output is fully supported by the nodes in the retrieval context. Great job on ensuring complete alignment!
CustomContextualRelevancy: The expected answer is fully supported by the retrieval context, which states that rights implied by admin cannot be denied, aligning with the expected output.
Correctness: The actual output provides detailed information on managing script rights, but it incorrectly suggests that script rights can be denied to a space administrator, contradicting the expected output that rights implied by admin right cannot be denied.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.7658054043883625
Individual_scores:
AnswerRelevancy: 0.8518518518518519
Faithfulness: 0.875
ContextualPrecision: 0.8541666666666666
ContextualRecall: 0.6666666666666666
CustomContextualRelevancy: 0.5
Correctness: 0.8471472411449898
Reasons:
AnswerRelevancy: The score is 0.85 because the response largely addresses the question about configuring permissions in XWiki, but includes several irrelevant links that do not directly assist with the specific configuration task described.

Faithfulness: The score is 0.88 because the claim that 'View' permission will still be inherited from the wiki-level settings contradicts the retrieval context, which states that page-level permissions override wiki-wide permissions.
ContextualPrecision: The score is 0.85 because most relevant nodes are ranked higher than irrelevant ones. The first node discusses 'rights inheritance and administration features' and 'inheritance of access rights,' which are relevant to setting view and edit permissions in XWiki. The second node explains the 'Nested Spaces' concept and rights management, also relevant to configuring space-level rights for teams. However, the third node, which is about string formatting and user identification in OpenID Connect, is not relevant to setting view/edit permissions in XWiki and should be ranked lower. Similarly, the fifth node, which describes the emulation of spaces and pages in XWiki, is not directly related to configuring view/edit permissions for teams and should be ranked lower than the relevant nodes.
ContextualRecall: The score is 0.67 because while several key concepts such as creating groups, setting wiki-wide and space-level rights, and inheriting rights for child pages are supported by nodes in the retrieval context, some introductory and summarizing sentences, as well as the reminder to adjust rights, are not directly referenced in the nodes.
CustomContextualRelevancy: The retrieval context mentions setting rights at different levels and inheritance, but lacks specific details about creating groups for each team and setting rights specifically for XWikiAllGroup or individual team spaces.
Correctness: The Actual Output includes all critical details such as setting global view rights, space-specific edit rights, and inheriting rights for child pages. It lacks explicit mention of creating groups for each team, but implies it in the instructions.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.7650360667931605
Individual_scores:
AnswerRelevancy: 0.7619047619047619
Faithfulness: 1.0
ContextualPrecision: 0.5
ContextualRecall: 0.7142857142857143
CustomContextualRelevancy: 0.8
Correctness: 0.8140259245684864
Reasons:
AnswerRelevancy: The score is 0.76 because the output mostly addresses the question about authenticating users with access tokens, but it includes multiple source links and a general introductory phrase that do not directly contribute to answering the question, thus preventing a higher score.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.50 because some relevant nodes are ranked lower than irrelevant ones. The first node is about 'LDAP configuration' and not about 'token-based authentication', yet it is ranked higher than the second node, which is relevant as it 'discusses the configuration of authorized applications'. Similarly, the third node focuses on 'OpenID Connect scopes', which is not directly related, but is ranked higher than the fourth node that 'describes the JWT token authenticator', which is directly relevant. This mix of rankings results in a balanced score.
ContextualRecall: The score is 0.71 because while several key aspects of the expected output, such as the use of JWTs and configuration steps, are supported by nodes in the retrieval context, there are missing details regarding generating a JWT token with specific claims and including the token in the 'Authorization' header, which are crucial for a complete understanding.

CustomContextualRelevancy: The retrieval context includes information about enabling the token authenticator, configuring authorized applications, and using JWTs, which aligns with the expected answer. However, details about specific JWT claims and the use of Ed25519 keys are not explicitly mentioned.
Correctness: The Actual Output correctly identifies the use of JWT for authentication, the need to set `xwiki.authentication.authclass`, and configuring 'Authorized Applications'. However, it inaccurately mentions a fallback authenticator which isn't in the Expected Output, and misses details on specific claims required in the JWT.
Question_language: en
Expected_answer_language: en
Answer_language: en

Model: AI.Models.qa_command-r_35B_Q4

File: qa_001_result.json

Overall_score: 0.30532534932314787
Individual_scores:
AnswerRelevancy: 0.8
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 0.0
CustomContextualRelevancy: 0.0
Correctness: 0.03195209593888694
Reasons:
AnswerRelevancy: The score is 0.80 because the response mostly addresses the issue of fixing the 'Bell' icon in XWiki, but includes irrelevant links that do not directly help solve the problem.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input question about enabling notifications in XWiki. The first node is about 'installation instructions and prerequisites for extensions', which does not help with the issue. The second node describes 'the functionality of the Alert menu and notifications' without addressing the enabling process. The third node lists 'dependencies related to the notifications extension', but again, does not solve the problem. Finally, the fourth node discusses 'overriding notification templates', which is unrelated to the query about enabling notifications.
ContextualRecall: The score is 0.00 because none of the sentences in the expected output can be linked to any nodes in the retrieval context, as there is no mention of enabling notifications or the `xwiki.properties` settings.
CustomContextualRelevancy: The retrieval context does not mention enabling notifications via the `notifications.enabled` setting in `xwiki.properties`, which is the key information in the expected output.
Correctness: The actual output does not address enabling notifications via 'notifications.enabled' in 'xwiki.properties' as stated in the expected output. Instead, it discusses issues with the 'Bell' icon and suggests installing the 'Alerts Application', which is unrelated to the expected task.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.6321514787801626
Individual_scores:

AnswerRelevancy: 0.8823529411764706
Faithfulness: 0.9090909090909091
ContextualPrecision: 0.75
ContextualRecall: 0.5
CustomContextualRelevancy: 0.5
Correctness: 0.251465022413596
Reasons:
AnswerRelevancy: The score is 0.88 because the response mostly addresses the issue of not receiving notifications in XWiki, but includes irrelevant source links that do not directly contribute to solving the problem.
Faithfulness: The score is 0.91 because the actual output incorrectly claims that no application types are enabled for notifications in XWiki versions less than 15.5, while the retrieval context specifies that certain types, such as Mention, are indeed enabled.
ContextualPrecision: The score is 0.75 because the first and fourth nodes in the retrieval context are relevant, explaining how to follow a user and listing ways to receive notifications, which directly address the issue of setting up notifications in XWiki. However, the second node, ranked higher than the fourth, discusses filters and settings without directly addressing enabling notifications, and the third node, also ranked higher than the fourth, mentions RSS feeds without providing information on enabling notifications. Additionally, the fifth node discusses the 'What's New' application, which is unrelated to managing notifications, contributing to the score not being higher.
ContextualRecall: The score is 0.50 because while some details about receiving notifications and following users are supported by nodes in the retrieval context, key instructions such as enabling notifications in `xwiki.properties` and subscribing to pages are missing from the context.
CustomContextualRelevancy: The retrieval context includes information about following users and receiving notifications, but lacks details on enabling notifications via `xwiki.properties` and watching pages or wikis.
Correctness: The actual output covers enabling notifications but includes unrelated details about filters, browser settings, and version discrepancies not present in the expected output. It lacks specific instructions on watching pages or users.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.9348777766659011
Individual_scores:
AnswerRelevancy: 0.75
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.859266659995407
Reasons:
AnswerRelevancy: The score is 0.75 because the response mostly addresses the question about enabling notifications for one's own actions in XWiki, but includes irrelevant source links that do not directly contribute to solving the issue.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Keep up the great work!
ContextualPrecision: The score is 1.00 because the relevant nodes in the retrieval context, which explain that 'By default, one don’t receive notification about one’s own activity' and provide solutions to 'disable this filter in the "Advanced filtering options" section of your own notification settings' (first node),

are ranked higher than irrelevant nodes. The irrelevant nodes, such as the one discussing 'administrators configuring notification settings for users' (third node), are correctly ranked lower. Great job on the precision!
ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context contains all the information from the expected answer, including the default behavior of not receiving notifications about one's own activity unless explicitly targeted, and the ability to disable this filter in the 'Advanced filtering options' section of the notification settings in the user profile.
Correctness: The actual output correctly states the default behavior and how to disable it, matching the expected output. It provides additional detail on navigating to the setting, which doesn't contradict the expected output but enhances user understanding.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.5644792278266976
Individual_scores:
AnswerRelevancy: 0.875
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 0.3
Correctness: 0.2118753669601857
Reasons:
AnswerRelevancy: The score is 0.88 because the response mostly addresses how to disable unwanted notifications in XWiki, but includes a statement about default notification settings, which is not directly relevant to the user's question.
Faithfulness: The score is 1.00 because there are no contradictions between the actual output and the retrieval context. Great job maintaining perfect alignment!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'getting a notification RSS feed and general notification settings' but does not address disabling notifications or automatic page watching. The second node talks about 'following a user and default watch settings' without mentioning how to disable notifications. The third node mentions 'filters for notifications' but lacks information on disabling notifications. The fourth node discusses 'enabling and disabling email notifications by an administrator,' which is not directly related to the user's concern. The fifth node is about 'administrators configuring notification settings and default auto-watched modes,' but it does not explain how a user can disable notifications. The sixth node is about 'overriding default notification templates and installation instructions,' which is not relevant to the user's query.
ContextualRecall: The score is 1.00 because every sentence in the expected output is fully supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context mentions default notification settings and user profile settings but lacks details on automatic page watching and deleting notification filters.
Correctness: The actual output does not mention automatic page watching after major modifications or how to change this setting in the profile, and introduces new details about following users and events, which are unsupported.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.799220159303947
Individual_scores:
AnswerRelevancy: 0.8333333333333334
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.4
CustomContextualRelevancy: 1.0
Correctness: 0.5619876224903482
Reasons:
AnswerRelevancy: The score is 0.83 because the answer mostly addresses the question about configuring XWiki for OpenID Connect, but includes some irrelevant details about LDAP Authenticator and skipping authentication, which are not directly related to enabling user accounts upon first login.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which mentions the property '# oidc.enableUser=true' directly related to enabling user accounts on first login, is ranked first. This ensures that the relevant information is prioritized over irrelevant nodes discussing topics like domain-based instance configuration, general OpenID Connect information, container authentication, and LDAP configuration.
ContextualRecall: The score is 0.40 because while the 1st node in retrieval context supports the mention of the `oidc.enableUser` property, it lacks explicit guidance on adding or updating properties, saving the file, and restarting the XWiki instance, as well as the outcome of automatically enabling user accounts upon first login.
CustomContextualRelevancy: The retrieval context contains the exact information found in the expected answer, specifically the property '# oidc.enableUser=true' which needs to be uncommented and set to 'true' in the xwiki.properties file.
Correctness: The actual output correctly mentions setting the `oidc.enableUser` to `true`, but it incorrectly refers to `xwiki.cfg` instead of `xwiki.properties` and adds unnecessary details about skipping authentication which aren't in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.8615847382953193
Individual_scores:
AnswerRelevancy: 0.9090909090909091
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.5
CustomContextualRelevancy: 1.0
Correctness: 0.7604175206810069
Reasons:
AnswerRelevancy: The score is 0.91 because the response effectively addresses the question about customizing user profile fields in XWiki with OpenID Connect, but includes a source link that doesn't directly contribute to the solution.
Faithfulness: The score is 1.00 because there are no contradictions; the actual output perfectly aligns with the retrieval context. Great job!

ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context is ranked first. It mentions 'It's possible to associate non standard properties coming from the OpenID Connect provider with the XWiki user' and provides examples of how to map properties using 'oidc.user.mapping', which directly relates to customizing user profile fields. Great job on getting it spot on!
ContextualRecall: The score is 0.50 because while the retrieval context provides information on mapping non-standard properties from the OpenID Connect provider to XWiki user properties (node 3), it lacks details on replacing placeholders with actual property names and the necessity of saving and restarting XWiki to apply changes.
CustomContextualRelevancy: The retrieval context contains all the information from the expected answer, including the ability to customize user profile fields using the 'oidc.user.mapping' property in the 'xwiki.properties' file, with examples provided.
Correctness: The actual output aligns with the expected output regarding the ability to customize user profile fields and the use of `oidc.user.mapping` in `xwiki.properties`. However, the example mappings differ; the expected uses `${oidc.user.subject}` while the actual uses `${oidc.user.given_name}`, `${oidc.user.family_name}`, and `${oidc.user.email}`. The sequence and logic are consistent, but the specific variable examples differ.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.7316034739687494
Individual_scores:
AnswerRelevancy: 0.8
Faithfulness: 0.6666666666666666
ContextualPrecision: 1.0
ContextualRecall: 0.45454545454545453
CustomContextualRelevancy: 0.9
Correctness: 0.5684087226003744
Reasons:
AnswerRelevancy: The score is 0.80 because the response mostly addresses the input question about configuring group synchronization between XWiki and an OpenID Connect provider. However, it includes some irrelevant statements that do not contribute to the explanation, preventing a higher score.
Faithfulness: The score is 0.67 because the actual output incorrectly states that adding the claim xwiki_groups to oidc.userinfoclaims is the default setting, which is not specified in the retrieval context. Additionally, the actual output refers to the oidc.groups.separator parameter for parsing groups, which is not mentioned in the retrieval context.
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than the irrelevant ones. Great job on getting everything in the right order!
ContextualRecall: The score is 0.45 because while some details about configuring group synchronization, such as claims and separators, are supported by nodes in the retrieval context, many procedural steps and introductory statements in the expected output are not explicitly covered by the nodes in the retrieval context.
CustomContextualRelevancy: The retrieval context includes detailed instructions on configuring group synchronization using properties like oidc.groups.claim, oidc.groups.separator, and oidc.groups.mapping, matching the expected output. However, it lacks explicit mention of saving the file and restarting the XWiki instance.
Correctness: The actual output correctly mentions configuring `oidc.groups.mapping` and `oidc.groups.separator`, similar to the expected output, but lacks detail on `oidc.groups.claim` setup

and saving changes. It introduces `oidc.groups.allowed` and `oidc.groups.forbidden`, which are not in the expected output, indicating discrepancies.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.5370942935287676
Individual_scores:
AnswerRelevancy: 0.875
Faithfulness: 0.8
ContextualPrecision: 0.3333333333333333
ContextualRecall: 0.8888888888888888
CustomContextualRelevancy: 0.2
Correctness: 0.12534353895038342
Reasons:
AnswerRelevancy: The score is 0.88 because the response largely addresses the question about setting upload restrictions, but it includes a link to basic concepts of XWiki, which is not directly relevant to the specific query.
Faithfulness: The score is 0.80 because the actual output incorrectly claims that the documentation lacks details on setting a size limit for uploads, while the retrieval context clearly states that administrators can set limits on the maximum size of an attachment in XWiki.
ContextualPrecision: The score is 0.33 because the relevant node, ranked third, provides information about 'setting the maximum size of an attachment in XWikiPreferences', which directly addresses the input query. However, irrelevant nodes are ranked higher, such as the first node discussing 'linking to external files and backlinks indexation', and the second node offering 'general information about XWiki's structure and features', neither of which relate to configuring upload restrictions or size limits.
ContextualRecall: The score is 0.89 because while the 3rd node in the retrieval context provides detailed instructions for configuring mimetypes and setting maximum attachment size, the overall goal of ensuring only images or PDF files of max 10MB can be uploaded is not explicitly stated in the retrieval context.
CustomContextualRelevancy: The retrieval context mentions setting maximum attachment size and mimetype restrictions but lacks specific details on configuring mimetypes and setting size to 10MB as outlined in the expected output.
Correctness: The actual output lacks specific steps to configure mimetype restrictions and max attachment size as detailed in the expected output, and suggests consulting documentation without addressing the task directly.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.569527885265984
Individual_scores:
AnswerRelevancy: 0.9090909090909091
Faithfulness: 0.6666666666666666
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 0.8
Correctness: 0.041409735838327646

Reasons:
AnswerRelevancy: The score is 0.91 because the response is mostly relevant and provides useful information on the topic, but it includes a link about content organization and nested pages migration, which is not directly related to denying script rights to a space administrator.
Faithfulness: The score is 0.67 because the actual output incorrectly claims that the 'Script' right is allowed for all users at the main wiki level and that all users can execute scripts unless revoked, whereas the retrieval context clarifies that this right is not granted by default in XWiki 14.10+.
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses programming rights but does not address denying script rights to a space administrator. The second node lists various rights without mentioning the inability to deny script rights. The third node explains script rights but not their denial. The fourth node provides technical details unrelated to denying script rights. The fifth node discusses rights conversion during migration without mentioning script rights denial.
ContextualRecall: The score is 1.00 because the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context confirms that rights implied by admin cannot be denied, aligning with the expected output. However, it does not explicitly mention the lack of support for denying implied rights in XWiki, leading to a slight deduction.
Correctness: The actual output does not address the key fact that rights implied by admin right cannot be denied in XWiki. It focuses on script rights management instead, which is not relevant to the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.645520425590486
Individual_scores:
AnswerRelevancy: 0.8947368421052632
Faithfulness: 0.7142857142857143
ContextualPrecision: 0.29166666666666663
ContextualRecall: 0.875
CustomContextualRelevancy: 0.5
Correctness: 0.5974333304852715
Reasons:
AnswerRelevancy: The score is 0.89 because the answer mostly addresses the question about configuring permissions in XWiki, but it includes irrelevant lists of sources that do not directly help with the configuration task.
Faithfulness: The score is 0.71 because the actual output inaccurately claims that you can set rights for each group on their corresponding space and access 'Space Administration', while the retrieval context indicates that the concept of space has been removed from the XWiki UI.
ContextualPrecision: The score is 0.29 because the relevant nodes in the retrieval context, specifically the fourth node which 'provides information on XWiki's permission system, including the ability to set view and edit rights at different levels,' and the sixth node which 'outlines the rules for setting rights in XWiki, including wiki-wide and page-level permissions,' are ranked lower than irrelevant nodes. The first node, which 'discusses the historical organization of content in XWiki and the concept of Nested Pages,' and the second node, which 'explains the parent/child relationship and the introduction of Nested Pages,' are ranked higher despite not addressing the configuration of access rights for teams. This misordering of relevant and irrelevant nodes results in a lower score.
ContextualRecall: The score is 0.88 because most of the expected output sentences are supported by the nodes in the retrieval context, such as setting rights at different levels and creating groups for each

team. However, there is a lack of specific mention in the nodes regarding adjusting rights for additional teams or spaces, which slightly lowers the score.
CustomContextualRelevancy: The retrieval context mentions setting rights at different levels and inheritance, but lacks specific details about creating groups for each team and setting rights for XWikiAllGroup as mentioned in the expected output.
Correctness: The actual output aligns with expected output on creating groups and setting edit rights, but lacks clarity on setting view rights for 'XWikiAllGroup' and ensuring inheritance for child pages.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.6891510956721144
Individual_scores:
AnswerRelevancy: 0.7142857142857143
Faithfulness: 1.0
ContextualPrecision: 0.5
ContextualRecall: 0.7142857142857143
CustomContextualRelevancy: 0.7
Correctness: 0.5063351454612569
Reasons:
AnswerRelevancy: The score is 0.71 because the response partially addresses the question about authenticating users with access tokens, but includes multiple irrelevant statements providing source links that do not directly answer the query.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.50 because relevant nodes are not consistently ranked higher than irrelevant nodes. The first node discusses 'LDAP authentication settings, which is unrelated to token-based authentication with JWTs' and should be ranked lower. The second node is relevant as it 'provides information on configuring authorized applications, which is relevant to setting up token-based authentication.' The third node, which 'describes OpenID Connect scopes and response types, which do not directly relate to JWT token-based authentication for the LLM Application,' should also be ranked lower. The fourth node is relevant as it 'describes the JWT token authenticator, which is directly relevant to the expected output about using access tokens for authentication.' Finally, the fifth node 'provides a general overview of XWiki authentication mechanisms, but does not specifically address JWT token-based authentication,' and should be ranked lower.
ContextualRecall: The score is 0.71 because while the retrieval context nodes provide substantial information about JWT authentication, such as enabling the token authenticator and configuring authorized applications, they lack specific details on generating a JWT token with required claims and including it in the 'Authorization' header.
CustomContextualRelevancy: The retrieval context mentions the use of JWT for authentication and configuring authorized applications, aligning with the expected answer. However, it lacks details on specific claims required in the token and the use of Ed25519 key for signing.
Correctness: The actual output mentions JWT and how to enable it, aligning partly with the expected output. However, it introduces OpenID Connect, which is additional incorrect information not present in the expected output. It also lacks details on configuring authorized applications, generating JWT tokens with specific claims, and the limitation regarding user creation.
Question_language: en
Expected_answer_language: en
Answer_language: en

Model: AI.Models.qa_gemma2_9B_Q4

File: qa_001_result.json

Overall_score: 0.3153455734980692
Individual_scores:
AnswerRelevancy: 0.8823529411764706
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 0.0
CustomContextualRelevancy: 0.0
Correctness: 0.009720499811944549
Reasons:
AnswerRelevancy: The score is 0.88 because the response is mostly relevant and helpful, addressing the issue of the missing 'Bell' notification icon. However, it includes repeated irrelevant statements about a source link that do not directly contribute to solving the problem, preventing a perfect score.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input query. The first node discusses 'installation instructions and prerequisites for extensions,' which does not address enabling notifications in XWiki. The second node describes 'the functionality of the notifications and the bell icon' but lacks information on enabling them. The third node lists dependencies unrelated to enabling notifications. Lastly, the fourth node covers 'overriding notification templates,' which is not helpful for the issue at hand.
ContextualRecall: The score is 0.00 because the retrieval context does not provide any information related to enabling notifications in the wiki or the specific settings mentioned in the expected output.
CustomContextualRelevancy: The retrieval context does not mention enabling notifications via the `notifications.enabled` setting in `xwiki.properties`.
Correctness: The actual output does not address enabling notifications via the `notifications.enabled` setting in `xwiki.properties` as described in the expected output. It discusses troubleshooting the Notifications Application and potential issues, which is unrelated to the expected task.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.6861828360491643
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.75
ContextualRecall: 0.5
CustomContextualRelevancy: 0.6
Correctness: 0.26709701629498606
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions. Great job maintaining accuracy and alignment with the retrieval context!
ContextualPrecision: The score is 0.75 because the first node in the retrieval context is relevant, explaining how to follow a user to receive notifications, and the fourth node is also relevant, providing an overview of ways to receive notifications in XWiki. However, the second node, discussing filters and

settings without addressing enabling notifications, and the third node, focusing on RSS feeds and settings, are both irrelevant and ranked higher than the fourth relevant node. Additionally, the fifth node, which is about the 'What's New' application, is not directly related to enabling notifications.
ContextualRecall: The score is 0.50 because half of the sentences in the expected output are well-supported by nodes in the retrieval context, such as the details on using the 'alert' menu and following a user. However, key setup instructions like enabling notifications in `xwiki.properties` and subscribing to pages are missing from the retrieval context.
CustomContextualRelevancy: The retrieval context includes information on following users and watching pages, but lacks details on enabling notifications in xwiki.properties and watching entire wikis.
Correctness: The actual output mentions enabling notification types and checking filters, which partially aligns with enabling notifications in the expected output. However, it lacks details about setting notifications in `xwiki.properties`, watching pages or users, and using the 'alert' menu. No mention of the 'network' tab in user profiles is made.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.8575288779709581
Individual_scores:
AnswerRelevancy: 0.6666666666666666
Faithfulness: 0.6666666666666666
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.8118399344924155
Reasons:
AnswerRelevancy: The score is 0.67 because the response partially addresses how to enable notifications for one's own actions in XWiki, but includes irrelevant source links that do not directly contribute to solving the issue.
Faithfulness: The score is 0.67 because the actual output incorrectly claims that the 'Own event filter' is usually enabled by default, while the retrieval context indicates that users do not receive notifications for their own actions by default, suggesting that the filter is indeed enabled by default.
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than irrelevant nodes. The first and second nodes provide precise information about enabling notifications for one's own actions in XWiki, while the third, fourth, and fifth nodes, which focus on administrative settings, RSS feeds, and general notifications, are correctly ranked lower as they do not address the specific issue.
ContextualRecall: The score is 1.00 because all sentences in the expected output are fully supported by the nodes in the retrieval context, showcasing a perfect alignment. Great job!
CustomContextualRelevancy: The retrieval context fully contains the expected answer, mentioning that by default, notifications about one's own activity are not received unless explicitly targeted, and this can be changed in the 'Advanced filtering options' of the user's notification settings.
Correctness: Both outputs mention that by default, one does not receive notifications about their own activity and that this can be changed in the notification settings. However, the actual output includes additional information about the specific filter name 'Own event filter' not mentioned in the expected output, but this does not contradict the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.608561647961182
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 0.3
Correctness: 0.351369887767092
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question about disabling notifications in XWiki without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'getting a notification RSS feed and enabling/disabling notification types', but it does not address 'how to disable notifications for pages you are not interested in'. The second node talks about 'following a user and default watch settings', which does not provide information on 'how to disable notifications for unwanted pages'. Subsequent nodes similarly fail to address the specific need of disabling notifications for unwanted pages, focusing instead on unrelated topics such as 'filtering notifications for hidden pages', 'enabling/disabling email notifications', 'administrative settings for notifications', and 'overriding default notification templates'.
ContextualRecall: The score is 1.00 because all sentences in the expected output are fully supported by the nodes in the retrieval context, demonstrating a perfect match. Great job!
CustomContextualRelevancy: The retrieval context mentions default notification settings and user profile settings but lacks details on major modifications automatically adding pages to watchlist and deleting custom filters.
Correctness: The actual output provides a general guide to notification settings in XWiki but lacks specific mention of 'Automatic page watching' and 'Custom Filters' as stated in the expected output. It does not address the automatic addition of major modification pages to the watchlist or the deletion of filters created that way.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.8866038555018895
Individual_scores:
AnswerRelevancy: 0.9230769230769231
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.6
CustomContextualRelevancy: 0.9
Correctness: 0.896546209934414
Reasons:
AnswerRelevancy: The score is 0.92 because the response is highly relevant to configuring OpenID Connect in XWiki, with only a minor mention of LDAP authentication that slightly detracts from the focus.
Faithfulness: The score is 1.00 because there are no contradictions between the actual output and the retrieval context. Great job on maintaining accuracy!

ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which mentions '# oidc.enableUser=true', is ranked first. This directly relates to enabling user accounts on first login, as required in the input. The irrelevant nodes, discussing topics like domain-based authentication, general OpenID Connect information, container authentication, and LDAP configuration, are ranked lower, ensuring the most pertinent information is prioritized.
ContextualRecall: The score is 0.60 because while some sentences in the expected output are supported by the 1st node in the retrieval context, others lack direct support from any node, leading to a moderate recall score.
CustomContextualRelevancy: The retrieval context includes the commented property '# oidc.enableUser=true', matching the expected answer's requirement to uncomment and set it to true, but lacks explicit instructions to save and restart XWiki.
Correctness: The actual output aligns with the expected output, detailing the modification of the `oidc.enableUser` property and restarting XWiki. However, it does not mention uncommenting the property.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.8491670649988531
Individual_scores:
AnswerRelevancy: 0.75
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.6
CustomContextualRelevancy: 1.0
Correctness: 0.7450023899931193
Reasons:
AnswerRelevancy: The score is 0.75 because the answer partially addresses the question about customizing user profile fields in XWiki, but includes irrelevant statements like a source link and a reference to documentation that do not directly answer the question.
Faithfulness: The score is 1.00 because there are no contradictions. Great job on maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which mentions 'It's possible to associate non standard properties coming from the OpenID Connect provider with the XWiki user' and provides an example of mapping, is ranked first. This directly relates to customizing user profile fields, while the irrelevant nodes, such as those discussing 'scopes and response types for OpenID Connect' (second node), 'endpoints and HTTP headers for OpenID Connect' (third node), 'group synchronization and customization templates' (fourth node), and 'in-wiki configuration and multiple configurations through cookies' (fifth node), are ranked lower. Great job on getting the order spot on!
ContextualRecall: The score is 0.60 because while the first node in the retrieval context supports the customization of user profile fields and provides examples of mapping them, it lacks details on replacing placeholders with actual property names and the need to save and restart XWiki.
CustomContextualRelevancy: The retrieval context includes the information about customizing user profile fields using the `oidc.user.mapping` property in the `xwiki.properties` file, which matches the expected output.
Correctness: Both outputs describe using the `oidc.user.mapping` property in `xwiki.properties` for customizing user profile fields from OpenID Connect. However, the actual output uses a different example and omits the restart step mentioned in the expected output.
Question_language: en

Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.8252876045024932
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.5
CustomContextualRelevancy: 0.9
Correctness: 0.5517256270149582
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question about configuring group synchronization between XWiki and an OpenID Connect provider, without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than the irrelevant ones. The first three nodes provide detailed information on configuring the `oidc.groups.claim` and `oidc.groups.mapping` properties, which are crucial for group synchronization as required by the input. The irrelevant nodes, ranked fourth and fifth, focus on general configuration and authentication management, which do not specifically address the group synchronization process.
ContextualRecall: The score is 0.50 because while there is a good match for configuring properties like `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping` with nodes in the retrieval context, key steps such as locating and opening the `xwiki.properties` file, replacing placeholder values, and saving the file are not supported by the retrieval context.
CustomContextualRelevancy: The retrieval context covers key aspects of group synchronization such as setting oidc.groups.claim, oidc.groups.separator, and oidc.groups.mapping, but lacks explicit instructions for locating and editing the xwiki.properties file.
Correctness: The actual output mentions configuring group synchronization and restarting XWiki, aligning partially with the expected output. However, it lacks details on setting the `oidc.groups.claim` and `oidc.groups.separator` properties, and uses `oidc.userinfoclaims` instead, which isn't in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.6387772333636667
Individual_scores:
AnswerRelevancy: 0.9473684210526315
Faithfulness: 1.0
ContextualPrecision: 0.3333333333333333
ContextualRecall: 0.8
CustomContextualRelevancy: 0.2
Correctness: 0.5519616457960351
Reasons:
AnswerRelevancy: The score is 0.95 because the response is highly relevant and addresses the question about configuring attachments, but it includes a link to general information that might not

directly answer the specific query.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.33 because the relevant node in the retrieval context, which provides information on setting the maximum size of an attachment and mentions the configuration parameter in the XWikiPreferences document, is ranked third. However, irrelevant nodes, such as the first node discussing linking to external files and images, and the second node describing the structure of XWiki pages, are ranked higher. These nodes do not address configuring upload restrictions or size limits, affecting the precision score.
ContextualRecall: The score is 0.80 because most of the detailed steps in the expected output are well-supported by nodes in the retrieval context, particularly nodes 3 and 4, which cover configuring mimetypes and setting maximum attachment sizes. However, the introductory and explanatory sentences lack direct support from specific nodes in the retrieval context.
CustomContextualRelevancy: Retrieval context mentions setting maximum attachment size and mimetype restrictions but lacks specific details on setting size to 10MB and configuring mimetypes in xwiki.properties.
Correctness: The actual output correctly sets the maximum upload size to 10MB and uses a whitelist for allowed mimetypes, but it specifies explicit image types instead of `image/*`, potentially missing some subtypes.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.6510085734578082
Individual_scores:
AnswerRelevancy: 0.8
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.10605144074684911
Reasons:
AnswerRelevancy: The score is 0.80 because the main content effectively addresses how to deny script rights to a space administrator, but the inclusion of multiple source links that do not directly answer the question slightly detracts from the overall relevancy.
Faithfulness: The score is 1.00 because there are no contradictions. Everything aligns perfectly with the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses programming rights and their implications, but does not address the denial of script rights to a space administrator. The second node provides a list of rights and their properties, but does not mention the inability to deny script rights to a space administrator. The third node explains the script right and its default settings, but does not mention the inability to deny it to a space administrator. The fourth node is about checking access rights programmatically, which does not relate to denying script rights to a space administrator. The fifth node discusses converting rights during page migration, which is unrelated to denying script rights to a space administrator.
ContextualRecall: The score is 1.00 because every part of the expected output is fully supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The expected answer is fully supported by the retrieval context, which states that admin rights imply other rights and cannot be denied, aligning with the expected output.

Correctness: The actual output provides a method to deny script rights to a space administrator, contradicting the expected output that states rights implied by admin cannot be denied.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.753344863004887
Individual_scores:
AnswerRelevancy: 0.92
Faithfulness: 0.875
ContextualPrecision: 0.6666666666666666
ContextualRecall: 0.875
CustomContextualRelevancy: 0.6
Correctness: 0.5834025113626561
Reasons:
AnswerRelevancy: The score is 0.92 because the response is mostly relevant and addresses the question about configuring permissions in XWiki. However, it includes irrelevant source links that do not directly relate to the configuration of permissions, which slightly lowers the score.
Faithfulness: The score is 0.88 because the actual output incorrectly claims that XWiki uses a hierarchical permission system where higher-level permissions are inherited by lower levels, contradicting the retrieval context which states that page-level permissions can override wiki-wide permissions.
ContextualPrecision: The score is 0.67 because the first node in the retrieval context is relevant, discussing 'rights inheritance and administration features' and 'Inheritance of access rights', which are pertinent to configuring view and edit permissions in XWiki. However, the second node, ranked higher than some relevant nodes, is irrelevant as it focuses on the 'Nested Pages concept' and naming conflicts, not on permissions. Similarly, the third node is unrelated, addressing string formatting and user identifiers, which do not pertain to access control. The fourth node is relevant, detailing XWiki's permission system with 'View' and 'Edit' rights, but it is ranked lower than the irrelevant nodes. Lastly, the sixth node is relevant, explaining 'wiki wide rights, granular page level rights', and permission priorities, but it is also ranked lower than some irrelevant nodes.
ContextualRecall: The score is 0.88 because most of the expected output aligns well with the nodes in the retrieval context, such as creating groups, setting wiki-wide and space-level rights, and inheriting rights for child pages. However, there is no specific mention in the nodes about adjusting rights for additional teams or spaces, which slightly affects the score.
CustomContextualRelevancy: The retrieval context mentions setting rights at different levels and inheriting rights for child pages, aligning with steps 2 and 4 of the expected output. However, it lacks specific instructions on creating groups for each team and setting space-level rights, which are crucial parts of the expected answer.
Correctness: The actual output covers group creation and space-level permissions, aligning with expected steps 1 and 3. However, it lacks explicit mention of wiki-wide rights and inheritance for child pages, and includes additional details on explicit denials and an admin guide not present in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.6296716824534421

Individual_scores:
AnswerRelevancy: 0.3333333333333333
Faithfulness: 1.0
ContextualPrecision: 0.5
ContextualRecall: 0.8
CustomContextualRelevancy: 0.8
Correctness: 0.3446967613873188
Reasons:
AnswerRelevancy: The score is 0.33 because the output contains multiple irrelevant statements that are source URLs, which do not address the question about authenticating users with access tokens. This significantly lowers the relevancy, but the score is not zero, indicating there might be some relevant content present.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.50 because the relevant nodes in the retrieval context are ranked lower than some irrelevant nodes. The first node is about 'LDAP configuration and does not mention token-based authentication or JWTs', which should be ranked lower than the second node that discusses 'Authorized Applications Configuration', relevant to token-based authentication. Similarly, the third node about 'OpenID Connect and OAuth scopes' should be ranked lower than the fourth node, which provides a description of 'authentication using signed JSON Web Tokens (JWT)', directly relevant to the input.
ContextualRecall: The score is 0.80 because most of the expected output sentences are supported by nodes in the retrieval context, particularly node 4 and node 2. However, there is one sentence that lacks direct support from the retrieval context, which slightly lowers the score.
CustomContextualRelevancy: The retrieval context includes information about enabling the token authenticator, configuring authorized applications, and using JWTs, which aligns with the expected answer. However, specific details about JWT claims and the use of Ed25519 keys are not explicitly mentioned.
Correctness: The actual output mentions the use of JWT for authentication but lacks details such as the required claims, Ed25519 signing, and configuration steps mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

Model: AI.Models.qa_llama3_1_402b

File: qa_001_result.json

Overall_score: 0.32111482221107307
Individual_scores:
AnswerRelevancy: 0.8947368421052632
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 0.0
CustomContextualRelevancy: 0.0
Correctness: 0.03195209116117512
Reasons:
AnswerRelevancy: The score is 0.89 because the response is mostly relevant and addresses the issue with the 'Bell' icon in XWiki, but the inclusion of source links that do not directly contribute to solving the problem slightly detracts from the overall relevance.
Faithfulness: The score is 1.00 because there are no contradictions. Great job on maintaining perfect alignment with the retrieval context!

ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'installation instructions and prerequisites', the second node describes 'how notifications are displayed', the third node lists 'dependencies for the notifications extension', and the fourth node talks about 'overriding notification templates'. None of these address enabling notifications in the wiki settings.
ContextualRecall: The score is 0.00 because the retrieval context does not contain any information related to enabling notifications in the wiki or the specific settings in `xwiki.properties`, making it impossible to support the expected output.
CustomContextualRelevancy: The retrieval context does not mention enabling notifications via the `notifications.enabled` setting in `xwiki.properties`.
Correctness: The actual output does not address enabling notifications via `xwiki.properties` as the expected output specifies. It focuses on installing the Notifications Application, which is not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.7181391987443496
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.75
ContextualRecall: 0.5
CustomContextualRelevancy: 0.6
Correctness: 0.45883519246609705
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the query about not receiving notifications in XWiki without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.75 because the first node in the retrieval context is relevant, providing information on how to follow a user to receive notifications, which is part of the expected output. However, the second node discusses filters and settings related to notifications but does not directly address enabling notifications or subscribing to pages or users, and should be ranked lower. Similarly, the third node talks about getting a notification RSS feed and settings, which does not provide information on enabling notifications or subscribing to specific content, and should also be ranked lower. The fourth node is relevant as it explains different ways to be notified in XWiki, aligning with the expected output. Lastly, the fifth node is about the 'What's New' application for news updates, which is not directly related to enabling notifications or subscribing to content in XWiki, and should be ranked lower.
ContextualRecall: The score is 0.50 because while some sentences, such as those about using the 'alert' menu (supported by node 4 in retrieval context) and following users (supported by node 1 in retrieval context), are well-supported, key details like enabling notifications in `xwiki.properties` and subscribing to pages or users are not covered by any nodes in the retrieval context.
CustomContextualRelevancy: The retrieval context covers following users and notification settings but lacks specific details on enabling notifications in xwiki.properties and watching pages.
Correctness: The actual output partially aligns with the expected output by mentioning the need to watch pages or users for notifications. However, it lacks specific instructions about enabling notifications in 'xwiki.properties' and does not cover all methods of watching locations and users described in the expected output. It introduces additional troubleshooting steps not found in the

expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.9761317181531796
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.8567903089190769
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question without any irrelevant information. Great job on maintaining focus and clarity!
Faithfulness: The score is 1.00 because there are no contradictions between the actual output and the retrieval context. Great job on maintaining perfect alignment!
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than irrelevant nodes. The first and second nodes directly address how to enable notifications for one's own actions in XWiki, while the third, fourth, and fifth nodes, which discuss unrelated topics such as administrator configurations, RSS feeds, and general notification methods, are correctly ranked lower. Great job on achieving perfect precision!
ContextualRecall: The score is 1.00 because every part of the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context contains all the information from the expected output, including the default behavior of not receiving notifications about one's own activity unless explicitly targeted, and the ability to change this setting in the 'Advanced filtering options' of the user profile.
Correctness: The actual output accurately describes the default behavior of not receiving notifications for one's own activity and the option to disable this filter under advanced settings, matching the expected output. However, it includes additional steps and version specificity not present in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.5463411898849923
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 0.75
CustomContextualRelevancy: 0.3
Correctness: 0.228047139309954
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question about disabling unwanted notifications in XWiki without any irrelevant information. Great job!

Faithfulness: The score is 1.00 because there are no contradictions. Great job on ensuring complete alignment with the retrieval context!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input query. The first node discusses 'getting a notification RSS feed and enabling/disabling notification types', but does not address how to disable notifications or change settings related to automatic page watching. The second node focuses on 'following a user and default watch settings', which also fails to provide information on disabling notifications or changing automatic page watching settings. Subsequent nodes similarly do not address the query, discussing topics like 'filtering notifications', 'enabling/disabling email notifications', 'administrators configuring notification settings', and 'overriding default notification templates', none of which are relevant to the user's question about disabling notifications in XWiki.
ContextualRecall: The score is 0.75 because while most of the expected output is supported by nodes in the retrieval context, such as the default settings for notifications and the ability to change settings, there is no explicit mention of deleting existing notification filters in the 'Custom Filters' list.
CustomContextualRelevancy: The retrieval context mentions default watch settings and user profile settings but lacks details on automatic page watching and deleting notification filters.
Correctness: The actual output focuses on disabling notifications and unwatching pages, lacking key details like automatic addition of modified pages to watch list, changing 'Automatic page watching' settings, and deleting filters from 'Custom Filters' list.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.824847605665772
Individual_scores:
AnswerRelevancy: 0.9
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.3333333333333333
CustomContextualRelevancy: 0.9
Correctness: 0.815752300661298
Reasons:
AnswerRelevancy: The score is 0.90 because the response is mostly relevant and addresses the question about configuring XWiki for OpenID Connect. However, it includes an irrelevant statement about LDAP authentication, which slightly detracts from its focus.
Faithfulness: The score is 1.00 because there are no contradictions. Great job on maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which mentions the property '# oidc.enableUser=true', is ranked first. This directly addresses the need to enable user accounts on first login using OpenID Connect, as required by the input. The irrelevant nodes, discussing topics like domain-based authentication, general OpenID Connect information, container authentication, and LDAP, are correctly ranked lower.
ContextualRecall: The score is 0.33 because only the first sentence regarding the property '# oidc.enableUser=true' is supported by the 1st node in the retrieval context, while the rest of the instructions and explanations lack direct support from any node.
CustomContextualRelevancy: The retrieval context includes the key information about the 'oidc.enableUser' property, which matches the expected answer's requirement to uncomment and set it to 'true'. However, it lacks explicit instructions on saving the file and restarting XWiki.
Correctness: The actual output correctly identifies the `oidc.enableUser` property in `xwiki.properties`, specifies setting it to `true`, and explains its purpose. However, it omits the instruction to uncomment

the property and the need to save the file and restart XWiki. It also contains additional information about other authentication methods not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.8888232496386848
Individual_scores:
AnswerRelevancy: 0.9166666666666666
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.6666666666666666
CustomContextualRelevancy: 1.0
Correctness: 0.7496061644987756
Reasons:
AnswerRelevancy: The score is 0.92 because the response is highly relevant and mostly addresses the question about customizing user profile fields in XWiki using OpenID Connect. However, it includes a source link that does not directly contribute to answering the specific query, slightly lowering the score.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because all relevant nodes are ranked higher than irrelevant nodes. The first node in the retrieval context is directly related to mapping additional user profile fields from the OpenID Connect provider to XWiki user properties, while the subsequent nodes discuss topics such as access requests for various claims, OpenID Connect provider endpoints, group synchronization, and in-wiki configuration, which are not directly related to the input query.
ContextualRecall: The score is 0.67 because while the main concept of mapping user profile fields from the OpenID Connect provider is supported by the 1st node in the retrieval context, specific instructions about replacing variables and restarting XWiki are not directly supported by any node in the retrieval context.
CustomContextualRelevancy: The retrieval context contains all the necessary information found in the expected answer, including the ability to customize user profile fields using the oidc.user.mapping property in the xwiki.properties file, with examples provided.
Correctness: The actual output accurately describes the customization of user profile fields using the oidc.user.mapping property, consistent with the expected output. However, it lacks details about restarting XWiki and does not use the same example properties and syntax provided in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.8462152928167339
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.5
CustomContextualRelevancy: 1.0
Correctness: 0.5772917569004032

Reasons:
AnswerRelevancy: The score is 1.00 because the response is perfectly relevant and directly addresses the question about configuring group synchronization between XWiki and an OpenID Connect provider. Great job!
Faithfulness: The score is 1.00 because there are no contradictions. Great job on maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than the irrelevant ones. The first three nodes provide detailed information on configuring properties like 'oidc.groups.claim', 'oidc.groups.separator', and 'oidc.groups.mapping', which are essential for group synchronization. The irrelevant nodes, ranked fourth and fifth, only offer general descriptions about OpenID Connect and authentication management, which do not directly aid in the configuration process.
ContextualRecall: The score is 0.50 because while some sentences in the expected output are supported by nodes in the retrieval context, such as the configuration examples and explanations (nodes 1 and 2), many sentences are general instructions or placeholders that do not directly match any specific part of the retrieval context.
CustomContextualRelevancy: The retrieval context contains all necessary details for configuring group synchronization, including setting the 'oidc.groups.claim', 'oidc.groups.separator', and 'oidc.groups.mapping' properties, which align with the expected output instructions.
Correctness: The actual output includes configuration steps for 'oidc.groups.claim' and 'oidc.groups.mapping' similar to the expected output but misses details on 'oidc.groups.separator'. It adds extra configurations like 'oidc.groups.allowed', 'oidc.groups.forbidden', and 'oidc.groups.prefix' not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.7046039616491253
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.3333333333333333
ContextualRecall: 0.6363636363636364
CustomContextualRelevancy: 0.5
Correctness: 0.7579268001977821
Reasons:
AnswerRelevancy: The score is 1.00 because the output perfectly addresses the input without any irrelevant statements. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.33 because the relevant node in the retrieval context, ranked third, provides information on setting maximum attachment size and mentions mimetype restrictions, which are directly related to the input. However, irrelevant nodes ranked first, second, fourth, and fifth discuss topics such as linking to external files and images, general overview of XWiki's structure, attachment display and security settings, and versioning and storage of attachments, none of which address the configuration of upload size or mimetype restrictions. These irrelevant nodes should be ranked lower than the relevant node.
ContextualRecall: The score is 0.64 because while nodes in the retrieval context support detailed steps for configuring mimetypes and maximum attachment size (as seen in sentences 2-7), introductory and explanatory sentences (1 and 8) and step titles (9 and 10) are not directly supported by the nodes.

CustomContextualRelevancy: The retrieval context mentions configuring maximum attachment size and mimetype restrictions but lacks specific details on setting 10MB limit and mimetype allowList.
Correctness: The actual output correctly mentions setting the maximum upload size to 10MB and configuring mimetype restrictions. However, it introduces the Attachment Validation Application not mentioned in the expected output and uses 'attachment.upload.allowedMimeTypes' instead of 'attachment.upload.allowList'.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.6651530325738324
Individual_scores:
AnswerRelevancy: 0.9375
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.05341819544299424
Reasons:
AnswerRelevancy: The score is 0.94 because the response is highly relevant to the question about denying script rights to a space administrator, with only a minor irrelevant mention of Nested Pages Migration.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses programming rights but does not address denying script rights to a space administrator. The second node outlines rights but does not mention the inability to deny script rights. The third node details the 'Script' right but does not mention denial to a space administrator. The fourth node talks about access checks, unrelated to denying script rights. The fifth node is about converting rights during migration, not related to denying script rights.
ContextualRecall: The score is 1.00 because the expected output is fully supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context confirms that rights implied by admin cannot be denied, matching the expected answer's statement about XWiki's lack of support for denying such rights.
Correctness: The actual output incorrectly states that script rights can be denied to a space administrator, while the expected output clearly states that rights implied by admin right cannot be denied in XWiki.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.7638220862121293
Individual_scores:
AnswerRelevancy: 0.8461538461538461
Faithfulness: 1.0
ContextualPrecision: 0.6666666666666666
ContextualRecall: 0.6666666666666666

CustomContextualRelevancy: 0.5
Correctness: 0.903445337785597
Reasons:
AnswerRelevancy: The score is 0.85 because the response largely addresses the question about configuring team access in XWiki, but the inclusion of multiple irrelevant source links slightly detracts from its overall relevance.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Keep up the great work!
ContextualPrecision: The score is 0.67 because the first node in the retrieval context is relevant as it discusses 'rights inheritance and administration features' related to setting view and edit permissions in XWiki. However, the second node, which is ranked higher than it should be, focuses on 'the parent/child relationship and the concept of Nested Pages,' which does not directly address configuring view/edit rights for teams. Additionally, the third node, about 'string formatting and user identification,' is unrelated to configuring access rights and should be ranked lower than the relevant nodes.
ContextualRecall: The score is 0.67 because while nodes in the retrieval context provide support for detailed steps such as creating groups, setting wiki-wide and space-level rights, and inheriting rights for child pages, introductory and summary statements in the expected output do not directly align with specific nodes in the retrieval context.
CustomContextualRelevancy: The retrieval context mentions rights inheritance and setting rights at different levels, which aligns partially with the expected answer's steps on setting wiki-wide and space-level rights. However, it lacks specific details about creating groups for each team and setting the 'Edit' right specifically for team spaces.
Correctness: The actual output closely matches the expected output with slight variations in wording. Both outline creating groups, setting view rights at the wiki level, setting edit rights at the space level, and ensuring rights inheritance, although 'XWikiAllGroup' is not explicitly mentioned in the actual output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.7128582034204819
Individual_scores:
AnswerRelevancy: 0.8260869565217391
Faithfulness: 1.0
ContextualPrecision: 0.5
ContextualRecall: 0.7142857142857143
CustomContextualRelevancy: 0.8
Correctness: 0.43677654971543767
Reasons:
AnswerRelevancy: The score is 0.83 because the main content effectively addresses the question about authenticating users with access tokens, but the presence of multiple irrelevant source link statements prevents the score from being higher.
Faithfulness: The score is 1.00 because there are no contradictions. Great job on maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 0.50 because the second node in the retrieval context is relevant, describing the configuration of 'Authorized Applications,' which pertains to using tokens for authentication, and the fourth node is also relevant, providing a description of the extension for authentication with JWTs. However, the first node, ranked highest, is irrelevant as it focuses on LDAP configuration without mentioning token-based authentication, and the third node discusses OpenID Connect scopes, which are not directly related to token-based authentication with JWTs. Additionally,

the fifth node talks about general authentication mechanisms without specifically addressing token-based authentication.
ContextualRecall: The score is 0.71 because while several aspects of the expected output are supported by the nodes in the retrieval context, such as the use of JWT for authentication and configuration of authorized applications, there are missing details like generating a JWT token with specific claims and including it in the 'Authorization' header.
CustomContextualRelevancy: The retrieval context includes information about enabling the token authenticator, configuring authorized applications, and using JWTs, which aligns with the expected answer. However, details about specific JWT claims and the use of Ed25519 keys are missing.
Correctness: The actual output mentions LLM Application Authenticator but lacks details on JWT claims, signing, and limitations noted in expected output. It also discusses OIDC and custom authenticators, which are irrelevant.
Question_language: en
Expected_answer_language: en
Answer_language: en

Model: AI.Models.qa_llama3_1_8b_Q4

File: qa_001_result.json

Overall_score: 0.3321727607743676
Individual_scores:
AnswerRelevancy: 0.8461538461538461
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 0.0
CustomContextualRelevancy: 0.0
Correctness: 0.14688271849235918
Reasons:
AnswerRelevancy: The score is 0.85 because the response largely addresses the issue of the Bell icon not being visible in XWiki, but it includes irrelevant source links that do not directly help in resolving the problem.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input question about enabling notifications in xwiki.properties. The first node discusses 'installation instructions and prerequisites', which do not relate to the issue at hand. The second node describes 'the functionality of the alert menu and notifications', but fails to address the enabling of notifications. The third node lists 'dependencies for the extension', which is unrelated to the problem. Finally, the fourth node talks about 'overriding notification templates', which is also not relevant. Therefore, no relevant nodes are ranked higher than irrelevant ones.
ContextualRecall: The score is 0.00 because the retrieval context does not provide any information about enabling notifications in the wiki or the specific settings in `xwiki.properties`.
CustomContextualRelevancy: The retrieval context does not mention enabling notifications via the `notifications.enabled` setting in `xwiki.properties`.
Correctness: The actual output does not mention enabling notifications in xwiki.properties, which is required in the expected output, focusing instead on troubleshooting missing Bell icon and dependencies.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.6255948333888044
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 0.6666666666666666
ContextualPrecision: 0.75
ContextualRecall: 0.5
CustomContextualRelevancy: 0.5
Correctness: 0.33690233366615974
Reasons:
AnswerRelevancy: The score is 1.00 because the response is perfectly relevant and directly addresses the issue of not receiving notifications in XWiki. Great job!
Faithfulness: The score is 0.67 because the actual output incorrectly claims that new users don't watch anything or enable notification types by default, whereas the retrieval context clarifies that in XWiki versions before 15.5, new users have the whole wiki watched by default, and in versions 15.5 and later, all application types are enabled by default.
ContextualPrecision: The score is 0.75 because the first and fourth nodes in the retrieval context are relevant, addressing how to receive notifications in XWiki, and are ranked higher. However, the second node, which discusses 'filters and settings related to notifications but does not directly address enabling notifications or subscribing to content', and the third node, which 'describes obtaining an RSS feed for notifications, which is not directly related to enabling notifications or subscribing to content', are ranked higher than the fourth relevant node. Additionally, the fifth node, which 'describes the 'What's New' application, which is not directly related to enabling notifications or subscribing to content', is ranked lower, which is correct.
ContextualRecall: The score is 0.50 because while the retrieval context covers aspects like enabling notifications through the 'alert' menu, following users, and using the 'network' tab (as seen in node 1), it lacks information about enabling notifications in `xwiki.properties`, subscribing to pages or users, and watching a page or its children.
CustomContextualRelevancy: The retrieval context covers user following and notification settings but lacks explicit mention of enabling notifications via `xwiki.properties` or watching pages and wikis.
Correctness: The actual output mentions watching users and enabling notifications but lacks details on enabling notifications in 'xwiki.properties' and the specific process of subscribing to pages or users. It also does not mention the 'alert' menu and 'network' tab.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.9455944490710917
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.6735666944265505
Reasons:
AnswerRelevancy: The score is 1.00 because the response is perfectly relevant and directly addresses the question about enabling notifications for one's own actions in XWiki. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!

ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than irrelevant nodes. The first node clearly states, 'By default, one don’t receive notification about one’s own activity... You have the ability to disable this filter in the "Advanced filtering options" section of your own notification settings,' which directly addresses the input. The second node also supports this by explaining how to enable notifications for one's own actions. Irrelevant nodes, such as the third node, which is about administrators configuring notifications, are ranked lower.
ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context contains all the information from the expected output, including the default setting of not receiving notifications about one's own activity unless explicitly targeted, and the ability to disable this filter in the 'Advanced filtering options' section of the user's notification settings.
Correctness: The actual output provides a method to enable notifications for one's own activity, aligning with the expected output's mention of this feature. However, it incorrectly states modifying the xwiki.properties file as an alternative, which is not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.5339119390513982
Individual_scores:
AnswerRelevancy: 0.9166666666666666
Faithfulness: 0.8
ContextualPrecision: 0.2
ContextualRecall: 0.75
CustomContextualRelevancy: 0.3
Correctness: 0.23680496764172276
Reasons:
AnswerRelevancy: The score is 0.92 because the answer effectively addresses the question about disabling notifications, but includes an unnecessary source link that doesn't directly pertain to the solution.
Faithfulness: The score is 0.80 because the actual output incorrectly suggests that users can disable unwanted notifications themselves, while the retrieval context specifies that only administrators have the ability to change users' notification settings in XWiki 13.2+.
ContextualPrecision: The score is 0.20 because the relevant node in the retrieval context, which is the fifth node, is ranked lower than several irrelevant nodes. The first node discusses 'getting a notification RSS feed and enabling/disabling notification types', which does not address disabling notifications for uninterested pages. The second node talks about 'following a user and default watch settings', which is not about disabling notifications for specific pages. The third node describes 'filtering notifications for hidden pages and minor events', which is also not relevant. The fourth node is about 'enabling/disabling email notifications and customizing email notification settings', which does not directly relate to the input query. The sixth node discusses 'overriding default notification templates and installation instructions', which is not relevant to the input. The relevant node, which should be ranked higher, mentions 'administrators can change notification settings and default auto-watched mode', directly relating to changing settings for automatic page watching.
ContextualRecall: The score is 0.75 because most aspects of the expected output are supported by nodes in the retrieval context, such as automatic page watching and notification settings. However, there is no specific mention of deleting existing notification filters in the 'Custom Filters' list, which affects the score.

CustomContextualRelevancy: The retrieval context mentions default notification settings and the ability to change them, but lacks details about automatic page watching and deleting filters from the 'Custom Filters' list.
Correctness: Actual output suggests disabling notifications via toggling settings, diverging from expected output, which focuses on 'Automatic page watching' and 'Custom Filters'.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.829918931861671
Individual_scores:
AnswerRelevancy: 0.9166666666666666
Faithfulness: 0.8333333333333334
ContextualPrecision: 1.0
ContextualRecall: 0.6
CustomContextualRelevancy: 0.9
Correctness: 0.7295135911700258
Reasons:
AnswerRelevancy: The score is 0.92 because the response is mostly relevant and helpful in addressing the OpenID Connect configuration issue, but it includes an unnecessary reference to LDAP authentication, which slightly detracts from its focus.
Faithfulness: The score is 0.83 because the actual output incorrectly states that OIDC authentication does not automatically enable user accounts upon their first login by default, whereas the retrieval context indicates that user accounts should indeed be enabled on first login by default in OIDC configuration.
ContextualPrecision: The score is 1.00 because the relevant node, which mentions the property 'oidc.enableUser=true' directly related to enabling user accounts on first login, is ranked first. Irrelevant nodes, such as those discussing domain-based authentication and session cookies (second node), general OpenID Connect information (third node), container authentication and Active Directory (fourth node), and LDAP configuration (fifth node), are ranked lower. Great job on the perfect ranking!
ContextualRecall: The score is 0.60 because while the retrieval context nodes provide information on defining and setting the 'oidc.enableUser' property (related to sentences 1, 2, and 3), they lack details on saving the file, restarting XWiki, and the automatic enabling of user accounts (related to sentences 4 and 5).
CustomContextualRelevancy: The retrieval context contains the relevant information about the 'oidc.enableUser' property, but lacks explicit instructions to uncomment and set it to 'true', save the file, and restart the XWiki instance.
Correctness: The actual output correctly mentions setting `oidc.enableUser` to `true` to enable accounts at first login, aligning with the expected output's key fact. However, it refers to modifying `/WEB-INF/xwiki.cfg` instead of `xwiki.properties`, and lacks mention of restarting the XWiki instance, which are discrepancies.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.921017685071977
Individual_scores:
AnswerRelevancy: 0.9285714285714286

Faithfulness: 0.8888888888888888
ContextualPrecision: 1.0
ContextualRecall: 0.8
CustomContextualRelevancy: 1.0
Correctness: 0.908645792971545
Reasons:
AnswerRelevancy: The score is 0.93 because the response is highly relevant to the input question about customizing user profile fields in XWiki using OpenID Connect. However, it includes a source reference that, while informative, does not directly address the customization query, slightly affecting the score.
Faithfulness: The score is 0.89 because the actual output claims a specific mapping method 'oidc.user.mapping=myxproperty1=${oidc.user.subject}' which is not mentioned in the retrieval context, indicating a slight misalignment.
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which includes information about mapping non-standard properties from the OpenID Connect provider to XWiki user properties, is ranked first. This directly addresses the input query about customizing user profile fields.
ContextualRecall: The score is 0.80 because the retrieval context supports most of the expected output, particularly the customization of user profile fields and mapping examples, as seen in the 1st node in retrieval context. However, it lacks explicit information about saving the file and restarting XWiki.
CustomContextualRelevancy: The retrieval context contains all information from the expected answer, including customization of user profile fields via the oidc.user.mapping property in xwiki.properties and examples of mapping custom properties.
Correctness: The actual output accurately reflects the expected output, covering customization of user profile fields using `oidc.user.mapping`. It provides examples similar to those in the expected output. The additional information about multiple mappings adds value without factual errors. The mention of `oidc.user.nameFormater` is extra but factually correct, causing a minor deduction.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.8271119936311105
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.46153846153846156
CustomContextualRelevancy: 0.9
Correctness: 0.6011335002482012
Reasons:
AnswerRelevancy: The score is 1.00 because the response is perfectly relevant and directly addresses the question about configuring group synchronization between XWiki and an OpenID Connect provider. Great job!
Faithfulness: The score is 1.00 because there are no contradictions. Great job maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than irrelevant nodes. The first three nodes provide detailed information about configuring the `oidc.groups.claim` and `oidc.groups.mapping` properties, which are essential for synchronizing groups between XWiki and an OpenID Connect provider. The irrelevant nodes, ranked fourth and fifth, only

offer general information about OpenID Connect and authentication management, which do not directly address group synchronization.
ContextualRecall: The score is 0.46 because while nodes in the retrieval context provide detailed information on configuring properties like `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping` (supporting sentences 3, 4, and 5), they lack instructions for file handling and specific replacements (unsupporting sentences 1, 2, 6, and 7).
CustomContextualRelevancy: The retrieval context contains detailed information about configuring group synchronization, including setting the `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping` properties, which aligns closely with the expected output. However, it lacks explicit instructions on locating and editing the `xwiki.properties` file and restarting the XWiki instance.
Correctness: The actual output includes key steps like setting `oidc.groups.claim` and using `oidc.groups.mapping`, but lacks specific instruction on opening and saving the `xwiki.properties` file, and omits the example claim and separator details.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.5688424645805416
Individual_scores:
AnswerRelevancy: 0.8333333333333334
Faithfulness: 0.8571428571428571
ContextualPrecision: 0.3333333333333333
ContextualRecall: 0.6363636363636364
CustomContextualRelevancy: 0.3
Correctness: 0.45288162731008985
Reasons:
AnswerRelevancy: The score is 0.83 because the response mostly addresses the question about restricting file types and sizes for uploads, but includes some irrelevant information about XWiki basic concepts and attachment versioning, which are not directly related to the specific requirement.
Faithfulness: The score is 0.86 because the actual output incorrectly suggests that attachment versioning can be disabled in XWiki by setting `xwiki.store.attachment.versioning=0`, whereas the retrieval context indicates that XWiki stores all document attachment versions by default and does not mention an option to disable this feature.
ContextualPrecision: The score is 0.33 because the relevant node in the retrieval context, which discusses 'setting the maximum size of an attachment in XWikiPreferences', is ranked third. However, irrelevant nodes are ranked higher, such as the first node that 'does not address configuring mimetypes or setting maximum upload sizes', and the second node that 'does not provide information on configuring upload restrictions or size limits'.
ContextualRecall: The score is 0.64 because while the 3rd node in the retrieval context provides detailed steps for configuring the maximum attachment size (supporting sentences in Step 2 of the expected output), it lacks information on configuring allowed mimetypes in the xwiki.properties file (Step 1), which is why the score isn't higher.
CustomContextualRelevancy: The retrieval context mentions setting maximum attachment size and mimetype restrictions but lacks specific details on configuring 'xwiki.properties' and setting size to 10MB.
Correctness: Actual output specifies the maximum upload size correctly but configures mimetypes incorrectly by including PostScript and specific image formats instead of a wildcard. It also suggests irrelevant steps like disabling versioning.
Question_language: en
Expected_answer_language: en

Answer_language: en

File: qa_009_result.json

Overall_score: 0.6254551924757624
Individual_scores:
AnswerRelevancy: 0.8
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 0.9
Correctness: 0.052731154854575066
Reasons:
AnswerRelevancy: The score is 0.80 because the response mostly addresses how to deny script rights to a space administrator, but includes an irrelevant statement about nested pages migration, which slightly detracts from its focus.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'programming rights and script execution' but doesn't address denying script rights to a space administrator. The second node provides a 'table of rights and their properties' without mentioning the denial of script rights. The third node explains 'script right and its default settings' but lacks information on denying these rights. The fourth node contains 'code snippets related to access checks' without addressing the denial of script rights. Lastly, the fifth node discusses 'converting rights and excluding pages' but is unrelated to denying script rights to a space administrator.
ContextualRecall: The score is 1.00 because the expected output is fully supported by the node in the retrieval context. Great job on achieving complete alignment!
CustomContextualRelevancy: The retrieval context confirms that rights implied by admin cannot be denied, aligning with the expected output. However, it does not explicitly state that XWiki does not support denying such rights.
Correctness: The actual output incorrectly states that script rights can be denied to a space administrator, which contradicts the expected output stating it is not supported in XWiki.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.7342815424108319
Individual_scores:
AnswerRelevancy: 0.7647058823529411
Faithfulness: 0.7142857142857143
ContextualPrecision: 0.6666666666666666
ContextualRecall: 0.875
CustomContextualRelevancy: 0.6
Correctness: 0.785030991159669
Reasons:
AnswerRelevancy: The score is 0.76 because the output partially addresses the configuration question by providing relevant information, but it includes multiple irrelevant source links that do not directly contribute to solving the problem.
Faithfulness: The score is 0.71 because the actual output incorrectly claims that 'View' rights can be added at the wiki level for all teams, while the retrieval context clarifies that such permissions at the

wiki-wide level are overridden by those set at a page level.
ContextualPrecision: The score is 0.67 because relevant nodes in the retrieval context are ranked higher than some irrelevant nodes, but not all. The first node discusses 'rights inheritance and administration features', which is relevant and ranked appropriately. However, the second node, which discusses 'parent/child relationship and nested spaces', is not directly related to configuring view and edit rights and should be ranked lower. Similarly, the third node about 'string formatting and user identification in OpenID Connect' is unrelated and should also be ranked lower. The fourth node correctly addresses the 'permission system in XWiki', making it relevant and well-placed. The fifth node, discussing 'creation of spaces and pages', is not directly related to the input and should be ranked lower. The sixth node provides information on 'setting wiki-wide and page-level rights', which is relevant and appropriately ranked.
ContextualRecall: The score is 0.88 because most of the expected output sentences are well-supported by the nodes in the retrieval context, such as the creation of groups and setting of rights at various levels. However, the lack of specific mention regarding the adjustment of rights for additional teams or spaces slightly reduces the score.
CustomContextualRelevancy: The retrieval context mentions setting rights at different levels and inheritance, aligning with the expected answer's steps on setting wiki-wide and space-level rights and inheriting rights for child pages. However, it lacks specific details on creating groups for each team and setting 'Edit' rights specifically for team groups, which are crucial steps in the expected answer.
Correctness: The actual output aligns with the expected output in setting view and edit rights, but lacks mention of 'Rights: Page & Children' for inheritance and refers to external links not present in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.6978200550989072
Individual_scores:
AnswerRelevancy: 0.6666666666666666
Faithfulness: 1.0
ContextualPrecision: 0.5
ContextualRecall: 0.7142857142857143
CustomContextualRelevancy: 0.8
Correctness: 0.505967949641062
Reasons:
AnswerRelevancy: The score is 0.67 because the response addresses the question about authenticating users with access tokens, but it includes multiple irrelevant source links that do not directly contribute to the answer.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.50 because relevant nodes in the retrieval context are mixed with irrelevant ones. For example, the first node discusses 'LDAP configuration, which is not relevant to token-based authentication,' yet it is ranked higher than the second node, which 'explains how to configure authorized applications, relevant to using access tokens for authentication.' Similarly, the third node about 'OpenID Connect scopes and response types' is not directly related to the input and is ranked higher than the fourth node, which 'describes the use of JSON Web Tokens for authentication,' directly relating to token-based authentication. This mix of relevant and irrelevant nodes affects the precision score.
ContextualRecall: The score is 0.71 because while several key aspects of the expected output, such as the use of JWT for authentication and the configuration of authorized applications, are supported by

nodes in the retrieval context, specific details like generating a JWT token with required claims and including it in the 'Authorization' header are not explicitly mentioned in any node.
CustomContextualRelevancy: The retrieval context confirms the use of JWT for authentication and mentions configuring authorized applications, matching the expected answer. However, it lacks specific details on JWT claims and token inclusion in the request header.
Correctness: The actual output correctly mentions configuring the xwiki.authentication.authclass property for token-based authentication and using JWT, but it omits details about the required claims in the token and Ed25519 signing key. It also lacks information on configuring authorized applications and the specific use of the 'Authorization' header.
Question_language: en
Expected_answer_language: en
Answer_language: en

Model: AI.Models.qa_mistral-nemo_12b_Q4

File: qa_001_result.json

Overall_score: 0.2779194661310235
Individual_scores:
AnswerRelevancy: 0.9259259259259259
Faithfulness: 0.6666666666666666
ContextualPrecision: 0
ContextualRecall: 0.0
CustomContextualRelevancy: 0.0
Correctness: 0.0749242041935482
Reasons:
AnswerRelevancy: The score is 0.93 because the response is highly relevant and addresses the issue with the bell icon in XWiki. However, it includes a source link that does not directly contribute to solving the problem, preventing a perfect score.
Faithfulness: The score is 0.67 because the actual output incorrectly states that the bell icon for notifications appears only if there is at least one unread notification, while the retrieval context does not specify this condition.
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input question. The first node discusses 'installation instructions and prerequisites for extensions,' which does not help with enabling notifications. The second node describes 'the functionality of the Alert menu,' but lacks information on enabling notifications. The third node lists 'dependencies related to the notifications extension,' yet fails to address enabling notifications. Lastly, the fourth node talks about 'overriding notification templates,' which is unrelated to the issue of enabling notifications.
ContextualRecall: The score is 0.00 because none of the sentences in the expected output are supported by any nodes in the retrieval context.
CustomContextualRelevancy: The retrieval context does not mention enabling notifications through the `xwiki.properties` file or the `notifications.enabled` setting.
Correctness: The actual output does not address enabling notifications via the `notifications.enabled` setting in `xwiki.properties`, which is crucial in the expected output. It instead focuses on troubleshooting the bell icon and dependencies.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.6566651881259623

Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 0.8
ContextualPrecision: 0.8055555555555555
ContextualRecall: 0.5
CustomContextualRelevancy: 0.5
Correctness: 0.33443557320021833
Reasons:
AnswerRelevancy: The score is 1.00 because the response is perfectly relevant and directly addresses the issue of not receiving notifications in XWiki. Great job!
Faithfulness: The score is 0.80 because the actual output incorrectly states that no application types are enabled for notifications by default in XWiki 15.5 and later, except specific ones like Mention, whereas the retrieval context clarifies that all application types are enabled by default. Additionally, the actual output inaccurately claims you can watch a user from their profile page in the 'network' tab, while the retrieval context specifies that the 'network' tab only shows the list of users you are watching, not a way to watch a user from there.
ContextualPrecision: The score is 0.81 because most relevant nodes in the retrieval context are ranked higher than irrelevant ones. For example, the first node explains 'how to follow a user to receive notifications', aligning well with the input query. However, the second node, which 'discusses filtering notifications and hidden pages', is not directly related to enabling notifications and should be ranked lower than the third node, which 'mentions the settings for notifications'. Similarly, the fifth node introduces 'the 'What's New' application', which is not directly related to the issue of not receiving notifications and should be ranked lower than the fourth node, which 'provides an overview of different ways to receive notifications in XWiki'.
ContextualRecall: The score is 0.50 because while some details like using the 'alert' menu and following users for notifications are supported by nodes in the retrieval context, key instructions such as enabling notifications via `xwiki.properties` and subscribing to pages or wikis are not covered.
CustomContextualRelevancy: The retrieval context includes information on following users and watching pages, but lacks details on enabling notifications via 'xwiki.properties' and the alert menu specifics.
Correctness: The actual output mentions that notifications are not enabled by default, which is not stated in the expected output. It correctly identifies the need to watch users or spaces and discusses notification settings, but adds information about filters and default settings that are not present in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.9790043232107569
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.8740259392645413
Reasons:
AnswerRelevancy: The score is 1.00 because the response is perfectly relevant and directly addresses the question about enabling notifications for one's own activity in XWiki. Great job!

Faithfulness: The score is 1.00 because there are no contradictions. Great job on maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than irrelevant nodes. The first node explains, 'By default, one don't receive notification about one's own activity... You have the ability to disable this filter in the "Advanced filtering options" section of your own notification settings,' which directly answers the input question. The second node also supports this by stating, 'By default, you won't receive notifications for actions done by yourself. This can be changed by switching off the Own event filter,' which is relevant to enabling notifications for one's own actions. The irrelevant nodes, such as the third node discussing 'administrator settings and notification preferences inheritance,' do not directly address the input question and are ranked lower. Great job maintaining perfect precision!
ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context fully contains the expected answer, detailing the default notification settings and the ability to change them in the 'Advanced filtering options' section.
Correctness: The actual output accurately reflects the expected output with minor discrepancies; it adds steps for disabling the filter, aligning with the expected logical flow and facts.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.5924714086474765
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 0.3
Correctness: 0.2548284518848593
Reasons:
AnswerRelevancy: The score is 1.00 because the response is perfectly relevant and directly addresses the question about disabling notifications in XWiki without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'how to get a notification RSS feed and general notification settings', which does not address disabling notifications or automatic page watching. The second node is about 'following a user and default watch settings', which also does not mention disabling notifications or adjusting automatic page watching settings. Subsequent nodes similarly fail to address the specific issue of disabling notifications or automatic page watching, focusing instead on unrelated aspects like 'filters for notifications', 'enabling or disabling email notifications', 'administrator settings', and 'overriding default notification templates'.
ContextualRecall: The score is 1.00 because all sentences in the expected output are fully supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context mentions default watch settings and user profile settings but lacks details on automatic page watching and deleting notification filters.
Correctness: The actual output discusses managing notification settings but lacks specifics on 'automatic page watching' and 'deleting custom filters' from the expected output.
Question_language: en
Expected_answer_language: en

Answer_language: en

File: qa_005_result.json

Overall_score: 0.8620039661838952
Individual_scores:
AnswerRelevancy: 0.9090909090909091
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.4
CustomContextualRelevancy: 1.0
Correctness: 0.8629328880124627
Reasons:
AnswerRelevancy: The score is 0.91 because the response is highly relevant and mostly addresses the question about configuring OpenID Connect in XWiki. However, it includes an irrelevant statement about the LDAP Authenticator, which slightly detracts from its relevance.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job maintaining accuracy!
ContextualPrecision: The score is 1.00 because the relevant node, which mentions the property '# oidc.enableUser=true' directly related to enabling user accounts on first login, is ranked first. This ensures that the most pertinent information is prioritized for the input query. Great job on achieving perfect precision!
ContextualRecall: The score is 0.40 because while the retrieval context (node 1) supports the mention of enabling user accounts on first login and the specific property 'oidc.enableUser=true', it lacks information on adding or updating properties, saving the file, restarting XWiki, and ensuring automatic account enablement.
CustomContextualRelevancy: The retrieval context contains the exact information needed to enable user accounts on first login by setting 'oidc.enableUser=true' in the xwiki.properties file, matching the expected output.
Correctness: The actual output correctly describes modifying the `oidc.enableUser` property in `xwiki.properties` to enable accounts on first login. However, it omits the step to uncomment the property, which is explicitly mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.8734007422660065
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.6
CustomContextualRelevancy: 1.0
Correctness: 0.6404044535960394
Reasons:
AnswerRelevancy: The score is 1.00 because the answer perfectly addresses the question without any irrelevant statements. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job maintaining accuracy!

ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context is ranked first, providing direct information on how to 'associate non standard properties coming from the OpenID Connect provider with the XWiki user', which is exactly what the input is asking about. The irrelevant nodes, which discuss topics such as 'access to various claims', 'configuration of OpenID Connect endpoints', 'group synchronization', and 'in-wiki configuration', are correctly ranked lower as they do not pertain to mapping user profile fields.
ContextualRecall: The score is 0.60 because while the retrieval context (node 3) supports the mapping example and customization process described in the expected output, it lacks explicit details on replacing 'myxproperty1' and 'myxproperty2' with actual XWiki user property names, and does not mention the need to save the 'xwiki.properties' file and restart XWiki.
CustomContextualRelevancy: The retrieval context contains all the information from the expected output, including mapping user profile fields using the `oidc.user.mapping` property in the `xwiki.properties` file and the example mappings provided.
Correctness: The actual output includes the necessary mapping example and explanation for `oidc.user.mapping`, aligning with the expected output. However, it introduces additional facts like `oidc.user.nameFormater`, group synchronization, and event listeners, which are not in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.8359659511269095
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.46153846153846156
CustomContextualRelevancy: 0.9
Correctness: 0.6542572452229954
Reasons:
AnswerRelevancy: The score is 1.00 because the output is perfectly relevant and directly addresses the question about configuring group synchronization between XWiki and an OpenID Connect provider. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than the irrelevant ones. The first three nodes provide detailed information on configuring properties like `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping`, which are crucial for group synchronization as required by the input. The irrelevant nodes, ranked fourth and fifth, discuss general OpenID Connect information and authentication management, which do not directly pertain to the group synchronization process.
ContextualRecall: The score is 0.46 because while some specific configuration details like `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping` are supported by nodes in the retrieval context, many introductory and procedural instructions such as locating and editing the `xwiki.properties` file, as well as saving changes and restarting XWiki, are missing from the retrieval context.
CustomContextualRelevancy: The retrieval context includes information about configuring group synchronization using the `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping` properties, which matches the expected output. However, it lacks explicit instructions on locating and editing the `xwiki.properties` file and restarting the XWiki instance.

Correctness: The actual output correctly addresses group synchronization setup, claim configuration, and group mapping, similar to the expected output. However, it introduces 'oidc.userinfoclaims' not present in the expected output and lacks the step to save and restart XWiki.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.6078383899888478
Individual_scores:
AnswerRelevancy: 0.9333333333333333
Faithfulness: 1.0
ContextualPrecision: 0.3333333333333333
ContextualRecall: 0.6363636363636364
CustomContextualRelevancy: 0.2
Correctness: 0.5440000369027834
Reasons:
AnswerRelevancy: The score is 0.93 because the response is highly relevant to the question about configuring attachment settings, but it includes a link to a general guide on XWiki basic concepts, which is not directly relevant to the specific question.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.33 because the relevant node in the retrieval context, which 'discusses setting limits on the maximum size of an attachment and provides steps to change it', is ranked third. This node should be ranked higher than the first node, which 'discusses linking to external files and images, which is unrelated to configuring upload restrictions or size limits', and the second node, which 'describes the structure of XWiki pages and their components, which does not relate to setting upload restrictions or size limits'.
ContextualRecall: The score is 0.64 because while several detailed instructions (sentences 2 to 9) are supported by nodes in the retrieval context, the introductory statements and explanations (sentences 1, 3, and 7) lack direct support from the nodes in the retrieval context.
CustomContextualRelevancy: The retrieval context mentions setting maximum attachment size and mimetype restrictions, but lacks specific details on configuring 'xwiki.properties' for mimetypes and setting size to 10MB.
Correctness: The actual output correctly outlines the steps to set size limit and mimetype restriction but incorrectly includes 'attachment.download.forceDownload' and 'attachment.upload.enableComments', which are not part of the expected output. It also omits the detailed steps for configuring the maximum attachment size in the XWiki user interface.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.627108410070746
Individual_scores:
AnswerRelevancy: 0.7142857142857143
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0

Correctness: 0.0483647461387619
Reasons:
AnswerRelevancy: The score is 0.71 because the answer partially addresses the question about denying script rights, but includes irrelevant source links and markdown format statements that do not directly contribute to the solution.
Faithfulness: The score is 1.00 because there are no contradictions. Great job on maintaining accuracy and alignment with the retrieval context!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses programming rights and script execution but does not address the denial of script rights to a space administrator. The second node lists rights and their default states but does not mention the inability to deny rights implied by admin rights. The third node explains the 'Script' right but does not mention the inability to deny rights implied by admin rights. The fourth node provides technical details about checking access rights but does not address the specific issue of denying script rights to a space administrator. The fifth node discusses converting rights during page migration, which is unrelated to denying script rights to a space administrator.
ContextualRecall: The score is 1.00 because every part of the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context confirms that rights implied by admin cannot be denied, matching the expected output.
Correctness: The actual output incorrectly suggests that the 'Script' right can be denied to a space administrator, contradicting the expected output that states implied rights by admin cannot be denied.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.6854949036976433
Individual_scores:
AnswerRelevancy: 0.8181818181818182
Faithfulness: 1.0
ContextualPrecision: 0.29166666666666663
ContextualRecall: 0.75
CustomContextualRelevancy: 0.3
Correctness: 0.9531209373373757
Reasons:
AnswerRelevancy: The score is 0.82 because the output mostly addresses the configuration question in XWiki, but it includes multiple source links that do not directly contribute to solving the problem, slightly reducing the relevancy.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.29 because the relevant nodes in the retrieval context are ranked lower than irrelevant nodes. The first node discusses 'content organization and rights inheritance in XWiki,' which does not provide specific steps for configuring permissions for teams, making it less relevant. The second node explains 'the parent/child relationship and the introduction of Nested Pages in XWiki,' which also does not address team access rights. The third node is about 'string formatting and OpenID Connect properties,' unrelated to the query. These irrelevant nodes are ranked higher than the fourth node, which provides relevant information about 'XWiki's permission system, including setting view and edit rights.' Additionally, the sixth node, which outlines 'basic rules for setting rights in XWiki,' is also relevant but ranked lower. This misranking of nodes contributes to the lower score.
ContextualRecall: The score is 0.75 because while most steps in the expected output align well with the nodes in retrieval context, such as setting rights and inheriting them for child pages, specific actions like

creating groups for each team and adjusting rights for additional teams or spaces are not directly supported by the nodes.
CustomContextualRelevancy: The retrieval context mentions rights inheritance and setting rights at different levels, but lacks specific details about creating groups for each team and setting rights specifically for 'XWikiAllGroup' or individual team spaces as described in the expected output.
Correctness: All factual statements in the actual output align with the expected output, including creating groups, assigning wiki-wide view rights, and setting space-specific edit rights with inheritance for child pages. No discrepancies or deviations are present.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.7158095865492499
Individual_scores:
AnswerRelevancy: 0.7142857142857143
Faithfulness: 1.0
ContextualPrecision: 0.5
ContextualRecall: 0.7142857142857143
CustomContextualRelevancy: 0.8
Correctness: 0.5662860907240704
Reasons:
AnswerRelevancy: The score is 0.71 because the response partially addresses the question about authenticating users with access tokens, but it includes multiple irrelevant source links that do not contribute to answering the question directly.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.50 because the second node in the retrieval context, which 'explains how to configure authorized applications', is relevant and ranked higher than some irrelevant nodes. However, the first node discusses 'LDAP configuration, which is unrelated to token-based authentication using JWTs', and should be ranked lower. Additionally, the third node 'describes various claims and scopes for access tokens, but does not directly address the process of authenticating users with JWTs', which should also be ranked lower than the relevant fourth node that 'provides a description of the extension that allows authentication using JWTs'.
ContextualRecall: The score is 0.71 because while nodes in the retrieval context provide information about authenticating users with JWTs, enabling the token authenticator, configuring authorized applications, and limitations on user creation, they lack details on generating a JWT token with specific signing key requirements and including the token in the 'Authorization' header.
CustomContextualRelevancy: The retrieval context includes details about configuring authorized applications and JWT usage, matching the expected answer's description of token-based authentication. However, it lacks specific details about JWT claims and the requirement for an Ed25519 key.
Correctness: The actual output correctly mentions using JWT for authentication, setting `xwiki.authentication.authclass`, and creating/updating users based on JWT. However, it introduces OpenID Connect, which is not part of the expected output, and lacks details about JWT claims, authorized applications, and token usage in headers.
Question_language: en
Expected_answer_language: en
Answer_language: en

Model: AI.Models.qa_mistral2_large

File: qa_001_result.json

Overall_score: 0.36422064452702657
Individual_scores:
AnswerRelevancy: 0.8709677419354839
Faithfulness: 0.8888888888888888
ContextualPrecision: 0.30952380952380953
ContextualRecall: 0.0
CustomContextualRelevancy: 0.0
Correctness: 0.11594342681397676
Reasons:
AnswerRelevancy: The score is 0.87 because the answer is mostly relevant to the question about the bell icon in XWiki, but it includes repeated irrelevant statements about a source link that do not directly address the issue.
Faithfulness: The score is 0.89 because the actual output incorrectly suggests that manual installation of the Notifications Application is possible when the Extension Manager is not functioning, contradicting the retrieval context which states that offline installation of extensions in XWiki is not supported.
ContextualPrecision: The score is 0.31 because relevant nodes in the retrieval context are ranked lower than irrelevant nodes. For example, the third node, which 'mentions various ways to receive notifications in XWiki and references the Notifications Application,' is relevant but ranked after nodes discussing 'installation instructions and prerequisites' (first node) and 'access issues and error messages related to configurable applications' (second node), which are not relevant to the issue of the bell icon not appearing. Additionally, the seventh node, which 'provides information on configuring notification settings in the administration,' is relevant but ranked after nodes that describe 'the content and behavior of the Alert menu' (fourth node) and list 'dependencies for a specific extension' (fifth node), which do not address the issue.
ContextualRecall: The score is 0.00 because none of the sentences in the expected output are supported by any nodes in the retrieval context.
CustomContextualRelevancy: The retrieval context does not mention the `notifications.enabled` setting or `xwiki.properties` file, which are crucial for enabling notifications as described in the expected output.
Correctness: The actual output does not mention enabling notifications in xwiki.properties, which is the key instruction in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.6978716227798297
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 0.875
ContextualPrecision: 1.0
ContextualRecall: 0.5
CustomContextualRelevancy: 0.5
Correctness: 0.3122297366789786
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question without any irrelevant information. Great job!
Faithfulness: The score is 0.88 because the actual output incorrectly claims that new users do not receive notifications until they start watching something and enable specific application types, whereas

the context clarifies that the whole wiki is watched by default for new users in versions before 15.5, with certain application types like Mention enabled.
ContextualPrecision: The score is 1.00 because the relevant node, which explains 'how to follow a user to receive notifications about their activities', is ranked first. This ensures that the most pertinent information is prioritized, providing an excellent retrieval result!
ContextualRecall: The score is 0.50 because while some sentences, like those about using the 'alert' menu (supported by node 6 in the retrieval context) and following a user (supported by node 1 in the retrieval context), are well-supported, critical details about enabling notifications in 'xwiki.properties' and subscribing to pages are missing in the retrieval context.
CustomContextualRelevancy: The retrieval context mentions following users and watching pages, which aligns with the expected output. However, it lacks details on enabling notifications via xwiki.properties and using the alert menu.
Correctness: The actual output provides reasons for not receiving notifications but misses key details like enabling notifications in xwiki.properties and subscribing to specific pages or users as mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.9514770440446026
Individual_scores:
AnswerRelevancy: 0.75
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.9588622642676154
Reasons:
AnswerRelevancy: The score is 0.75 because the response partially addresses how to enable notifications for one's own actions in XWiki, but includes irrelevant source links that do not directly contribute to solving the issue.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant nodes in the retrieval context, which explain how to enable notifications for one's own actions by mentioning options like 'Advanced filtering options' and 'Own event filter', are ranked higher than the irrelevant nodes. The irrelevant nodes discuss topics such as following a user, administrator settings, and RSS feeds, which do not pertain to the input query. Great job on getting it just right!
ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context fully includes the expected answer's information about not receiving notifications for one's own activity by default and the ability to disable this filter in the 'Advanced filtering options' section of the notification settings in the user profile.
Correctness: The actual output aligns with the expected output by correctly stating the default notification setting and the ability to disable the filter in the 'Advanced filtering options' of the notification settings in the user profile.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.4803671637507298
Individual_scores:
AnswerRelevancy: 0.9583333333333334
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 0.5
CustomContextualRelevancy: 0.2
Correctness: 0.22386964917104518
Reasons:
AnswerRelevancy: The score is 0.96 because the response effectively addresses how to disable notifications in XWiki, but includes a source link that doesn't directly answer the question, slightly affecting its relevance.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'following a user and default watch settings', which does not help in disabling notifications for pages you are not interested in. The second node talks about 'selecting the level of details in email notifications and macros for settings', which is not relevant to the problem of disabling notifications. The third node covers 'enabling and disabling the extension and email features', but does not address the specific need to disable notifications for certain pages. All nodes fail to address the input query, resulting in a score of 0.00.
ContextualRecall: The score is 0.50 because only half of the expected output is supported by the nodes in the retrieval context. Specifically, sentences 3 and 4 are supported by nodes 4 and 5 in the retrieval context, while sentences 1 and 2 lack direct support.
CustomContextualRelevancy: The retrieval context mentions default watch settings and user profile settings but lacks details on automatic page watching and deleting custom filters.
Correctness: The actual output focuses on managing notifications broadly and does not mention automatic page watching or custom filters. It introduces the xwiki.properties file and macros, which are not in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.9003313056717284
Individual_scores:
AnswerRelevancy: 0.9333333333333333
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.6
CustomContextualRelevancy: 1.0
Correctness: 0.8686545006970372
Reasons:
AnswerRelevancy: The score is 0.93 because the response is highly relevant to configuring OpenID Connect in XWiki, but it slightly deviates by mentioning LDAP Authenticator, which is unrelated to the user's query.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant node, which contains the property '# oidc.enableUser=true', is ranked first. This directly addresses the need to enable user accounts on first

login, ensuring that all irrelevant nodes, such as those discussing 'domain-based instance authentication and session cookie configuration' (second node), 'resetting passwords and different authentication mechanisms' (third node), and others, are ranked lower.
ContextualRecall: The score is 0.60 because while the nodes in the retrieval context support the instructions to enable user accounts by setting the `oidc.enableUser` property, they do not mention the need to save the file and restart XWiki, nor do they explain the outcome of enabling new user accounts.
CustomContextualRelevancy: The retrieval context contains the exact information found in the expected answer, specifically the property '# oidc.enableUser=true' in the 'xwiki.properties' file, which needs to be uncommented and set to 'true' to enable user accounts on first login via OpenID Connect.
Correctness: The actual output correctly describes setting the oidc.enableUser property to true in the xwiki.properties file and restarting XWiki. However, it lacks guidance on uncommenting the line and uses a different format for adding the property. It does not mention the comments about enabling user accounts, which are present in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.8863405953627065
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.625
ContextualRecall: 0.8
CustomContextualRelevancy: 1.0
Correctness: 0.8930435721762396
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.62 because the first node in the retrieval context is relevant and directly addresses customizing user profile fields by mentioning 'It's possible to associate non standard properties coming from the OpenID Connect provider with the XWiki user'. However, several irrelevant nodes are ranked higher than the second relevant node. For example, the second node discusses 'access token claims like profile, email, and address', which are unrelated to the task, and the third node focuses on 'group synchronization and custom claims for id tokens', which do not pertain to mapping user profile fields. These irrelevant nodes should be ranked lower than the eighth node, which is relevant as it lists 'available variables like oidc.user.subject and oidc.user.mail', crucial for mapping additional user profile fields.
ContextualRecall: The score is 0.80 because most of the expected output is supported by the nodes in the retrieval context, particularly regarding the customization and mapping of user profile fields. However, the instruction to save and restart XWiki is not found in the retrieval context, which affects the score.
CustomContextualRelevancy: The retrieval context includes detailed instructions on customizing user profile fields using the `oidc.user.mapping` property in the `xwiki.properties` file, matching the expected output.
Correctness: The actual output accurately describes customizing user profile fields using `oidc.user.mapping` in `xwiki.properties`, includes examples and the need to save configurations, similar to the expected output. However, it omits the specific instruction to restart XWiki to apply changes.

Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.7450739750657117
Individual_scores:
AnswerRelevancy: 0.9210526315789473
Faithfulness: 1.0
ContextualPrecision: 0.8333333333333334
ContextualRecall: 0.46153846153846156
CustomContextualRelevancy: 0.7
Correctness: 0.5545194239435276
Reasons:
AnswerRelevancy: The score is 0.92 because the response is mostly relevant and provides useful information on configuring group synchronization. However, the repeated and unnecessary 'For example:' statements slightly detract from the overall clarity and focus, preventing a perfect score.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.83 because the first two nodes in the retrieval context are relevant, discussing 'oidc.groups.mapping' and 'oidc.groups.separator', which are crucial for configuring group synchronization. However, the third node is less relevant as it focuses on a general description and dependencies of the OpenID Connect extension. The fourth node, discussing authentication management and endpoints, and the fifth node, listing dependencies, are also not directly related to the configuration process. The sixth node is relevant, discussing 'oidc.groups.claim', which is important for the configuration process.
ContextualRecall: The score is 0.46 because while several specific configurations like `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping` are supported by nodes in the retrieval context, many general instructions such as locating and editing the `xwiki.properties` file, and saving changes are not directly found in the context.
CustomContextualRelevancy: The retrieval context includes details on configuring group synchronization, such as `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping`, but lacks explicit instructions on locating and editing the `xwiki.properties` file and restarting the XWiki instance.
Correctness: The actual output includes most steps such as configuring the 'oidc.groups.claim' and 'oidc.groups.separator' properties, but introduces additional concepts like 'oidc.groups.prefix' and 'oidc.groups.allowed', which are not in the expected output. Also, it mentions adding the claim 'xwiki_groups' to 'oidc.userinfoclaims' which is not specified in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.6261110027809882
Individual_scores:
AnswerRelevancy: 0.9333333333333333
Faithfulness: 1.0
ContextualPrecision: 0.2
ContextualRecall: 0.6666666666666666
CustomContextualRelevancy: 0.3

Correctness: 0.6566660166859294
Reasons:
AnswerRelevancy: The score is 0.93 because the response is highly relevant and mostly addresses the question about restricting attachments to images or PDFs of a certain size. However, it includes URLs that are not directly related to configuring these specific attachment restrictions, which slightly affects the score.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.20 because the relevant node, which 'mentions setting limits on the maximum size of an attachment and provides steps to change it', is ranked fifth. This means it is ranked lower than several irrelevant nodes, such as the first node that 'discusses linking to external files and images, which is unrelated to configuring upload restrictions or size limits'.
ContextualRecall: The score is 0.67 because while the retrieval context (nodes) provides detailed steps for configuring the maximum attachment size (as seen in the 5th node), it lacks information on configuring allowed mimetypes and ensuring only images or PDF files can be uploaded, which are crucial parts of the expected output (as mentioned in sentences 1 and 2).
CustomContextualRelevancy: The retrieval context mentions setting maximum attachment size and mimetype restrictions but lacks specific details on configuring xwiki.properties and setting size to 10MB.
Correctness: The actual output accurately describes the maximum size configuration, but uses a different MIME type configuration approach by listing specific subtypes rather than a wildcard and includes additional properties not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.6534252060280247
Individual_scores:
AnswerRelevancy: 0.8181818181818182
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.10236941798632948
Reasons:
AnswerRelevancy: The score is 0.82 because the response mostly addresses the question about denying script rights to a space administrator, but it includes multiple irrelevant source links that do not contribute to answering the question.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'programming rights and script execution but does not address denying script rights to a space administrator.' Similarly, the second node 'lists various rights and their default states, but it does not mention the inability to deny rights implied by admin rights.' The third node provides 'a tabular view of rights but does not mention the inability to deny script rights to a space administrator.' These irrelevant nodes are ranked higher than any relevant nodes, resulting in a score of 0.00.
ContextualRecall: The score is 1.00 because every part of the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context confirms that rights implied by admin cannot be denied, aligning with the expected output.

Correctness: The actual output suggests denying script rights at the space level, contradicting the expected output that implies rights implied by admin rights cannot be denied in XWiki.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.5975347035283416
Individual_scores:
AnswerRelevancy: 0.8571428571428571
Faithfulness: 1.0
ContextualPrecision: 0.20833333333333331
ContextualRecall: 0.5
CustomContextualRelevancy: 0.3
Correctness: 0.7197320306938592
Reasons:
AnswerRelevancy: The score is 0.86 because the answer mostly addresses the configuration question in XWiki, but it includes multiple source links that do not directly contribute to solving the problem, which slightly detracts from its relevancy.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.21 because the relevant nodes in the retrieval context, which provide information on XWiki's permission system and basic rules for setting rights, are ranked lower (sixth and eighth) than several irrelevant nodes. These irrelevant nodes, such as the first node discussing the organization of content and the second node focusing on parent/child relationships, do not address access control or rights management and should be ranked lower.
ContextualRecall: The score is 0.50 because while the retrieval context (node 6) supports the general idea of setting wiki-wide and page-level rights (related to sentences 2, 3, and 4), it lacks specific details about creating groups for each team (sentence 1) and ensuring all teams can view all spaces and only edit their own space (concluding sentences).
CustomContextualRelevancy: The retrieval context mentions setting rights at different levels and inheritance of access rights, which partially aligns with setting wiki-wide and space-level rights. However, it lacks specific details on creating groups for each team and setting 'Edit' rights for team-specific spaces as outlined in the expected output.
Correctness: The actual output aligns with the expected output in terms of global view rights and specific edit rights for each team. However, it lacks creating groups for each team and setting wiki-wide rights for XWikiAllGroup. It includes denying edit rights for other teams, which is not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.7209701682174745
Individual_scores:
AnswerRelevancy: 0.7058823529411765
Faithfulness: 1.0
ContextualPrecision: 0.5
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0

Correctness: 0.11993865636367071
Reasons:
AnswerRelevancy: The score is 0.71 because the main content of the output addresses the question about authenticating users with access tokens, but the inclusion of multiple irrelevant source links prevents the score from being higher.
Faithfulness: The score is 1.00 because there are no contradictions, indicating a perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.50 because relevant nodes in the retrieval context, such as the second node discussing 'the configuration of authorized applications' and the fourth node detailing 'JWT token requirements and signing process', are ranked lower than irrelevant nodes. For instance, the first node, which is about 'LDAP configuration and does not mention token-based authentication or JWT', is ranked higher than these relevant nodes. This misordering affects the precision score, as irrelevant nodes are not ranked lower than relevant ones.
ContextualRecall: The score is 1.00 because every sentence in the expected output is fully supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context contains all the necessary information from the expected answer, including the use of JWT tokens with specific claims, Ed25519 key signing, configuration steps in 'xwiki.properties', and limitations on user creation and updates.
Correctness: The actual output describes the OpenID Connect Authenticator, which differs from the expected output's Token-based authentication for the LLM Application extension, including different setup details and token requirements.
Question_language: en
Expected_answer_language: en
Answer_language: en

Model: AI.Models.qa_mixtral-8x22b

File: qa_001_result.json

Overall_score: 0.32705001693834407
Individual_scores:
AnswerRelevancy: 0.7777777777777778
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 0.0
CustomContextualRelevancy: 0.0
Correctness: 0.18452232385228687
Reasons:
AnswerRelevancy: The score is 0.78 because the output contains useful information related to XWiki, but it includes multiple irrelevant statements that do not directly address the specific issue of the missing bell icon.
Faithfulness: The score is 1.00 because there are no contradictions, indicating a perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'installation instructions for extensions', which does not help with enabling notifications. The second node describes 'functionality of the Alert menu', not addressing the issue of enabling notifications. The third node lists 'dependencies', offering no information on enabling notifications. Finally, the fourth node talks about 'customizing notification templates', unrelated to the problem of enabling notifications.
ContextualRecall: The score is 0.00 because the expected output about enabling notifications and setting `notifications.enabled` in `xwiki.properties` is not supported by any information in the nodes in the retrieval context.

CustomContextualRelevancy: The retrieval context lacks any mention of enabling notifications via the 'notifications.enabled' setting in 'xwiki.properties'.
Correctness: The actual output provides troubleshooting steps for the Alerts and Notifications Applications, but does not mention enabling notifications in `xwiki.properties` as specified in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.6779018897754411
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.75
ContextualRecall: 0.5
CustomContextualRelevancy: 0.5
Correctness: 0.3174113386526465
Reasons:
AnswerRelevancy: The score is 1.00 because the response is perfectly relevant and directly addresses the question about not receiving notifications in XWiki. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.75 because the relevant nodes in the retrieval context are mostly ranked higher than the irrelevant ones. However, the second node, which discusses 'filters and technical settings that do not directly address the issue of not receiving notifications or how to enable them,' and the third node, which focuses on 'RSS feeds and notification macros, which are not directly related to enabling notifications or ensuring they are received,' are ranked higher than the fourth node. The fourth node outlines 'different ways to receive notifications in XWiki, including the Notifications Application and mentions,' which is relevant to the input, and should be ranked higher.
ContextualRecall: The score is 0.50 because half of the expected output sentences are supported by nodes in the retrieval context, such as the details about using the 'alert' menu and following users, while the other half, like enabling notifications in `xwiki.properties` and subscribing to pages, are not directly supported by any nodes.
CustomContextualRelevancy: The retrieval context includes information on following users and notification settings but lacks details on enabling notifications in xwiki.properties and watching pages.
Correctness: The actual output correctly advises enabling notifications settings but lacks instructions about subscribing to pages or users. It mentions default settings for new users and customization, which is additional information not found in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.9751172472608903
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 1.0

CustomContextualRelevancy: 1.0
Correctness: 0.8507034835653414
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions. Great job maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 1.00 because the relevant nodes in the retrieval context are perfectly ranked higher than the irrelevant nodes. The first and second nodes directly address the issue of not receiving notifications for one's own activity by mentioning options to disable filters in notification settings, while the third, fourth, and fifth nodes discuss unrelated topics such as administrator settings and RSS feeds, which do not pertain to the specific issue.
ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context contains all the information from the expected output, including the default behavior of not receiving notifications about one's own activity unless explicitly targeted, and the ability to disable this filter in the 'Advanced filtering options' section of the notification settings.
Correctness: The actual output correctly states that by default, you won't receive notifications for your own actions and describes how to change this setting. It mentions the 'Own Event Filter' and suggests it might be in 'Advanced filtering options,' aligning closely with the expected output. However, it lacks explicit mention of the user profile location for the settings.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.5775503012533487
Individual_scores:
AnswerRelevancy: 0.8666666666666667
Faithfulness: 0.875
ContextualPrecision: 0.2
ContextualRecall: 1.0
CustomContextualRelevancy: 0.4
Correctness: 0.12363514085342604
Reasons:
AnswerRelevancy: The score is 0.87 because the response effectively addresses how to disable notifications in XWiki, but includes irrelevant source links that do not directly answer the question.
Faithfulness: The score is 0.88 because the claim implies notifications are received only after changes to settings, whereas the context states notifications are received after enabling notification types.
ContextualPrecision: The score is 0.20 because the relevant node, ranked fifth, correctly addresses changing notification settings by mentioning that 'administrators can change the default auto-watched mode and users can change their notification settings.' However, it is ranked lower than four irrelevant nodes. For example, the first node discusses 'how to get an RSS feed of notifications,' which is unrelated to disabling notifications. Similarly, the second node talks about 'following a user and default watch settings,' which does not provide information on how to disable notifications.
ContextualRecall: The score is 1.00 because every sentence in the expected output is fully supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context mentions default watch settings and user profile settings but lacks details about automatic page watching and deleting notification filters.

Correctness: The actual output describes managing notifications in XWiki but lacks the specific details about automatic page watching, changing settings, and deleting filters described in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.9338159333409104
Individual_scores:
AnswerRelevancy: 0.9230769230769231
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.8
CustomContextualRelevancy: 0.9
Correctness: 0.9798186769685389
Reasons:
AnswerRelevancy: The score is 0.92 because the response is highly relevant and provides useful information about configuring OpenID Connect in XWiki. However, it includes an irrelevant statement about LDAP authentication, which slightly detracts from its focus on the specific OpenID Connect configuration issue.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which mentions the property 'oidc.enableUser=true', is ranked first. This directly addresses the configuration needed to enable user accounts on first login, ensuring that irrelevant nodes discussing topics like domain-based instances, cookie configurations, and LDAP are ranked lower.
ContextualRecall: The score is 0.80 because the retrieval context nodes cover most of the expected output, including enabling user accounts on first login and the relevant property settings, but they do not mention the need to save the file and restart the XWiki instance.
CustomContextualRelevancy: The retrieval context includes the commented line '# oidc.enableUser=true', which matches the expected answer's requirement to uncomment and set this property to true. However, the context does not explicitly mention saving the file and restarting XWiki.
Correctness: The actual output accurately describes adjusting the xwiki.properties file, including setting oidc.enableUser to true and restarting XWiki, aligning with the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.918798554870337
Individual_scores:
AnswerRelevancy: 0.9
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.8
CustomContextualRelevancy: 1.0
Correctness: 0.8127913292220222
Reasons:

AnswerRelevancy: The score is 0.90 because the response effectively addresses the main question about customizing user profile fields in XWiki using OpenID Connect, but includes a source link that is not directly relevant to the query.
Faithfulness: The score is 1.00 because there are no contradictions. Great job on maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which states 'It's possible to associate non standard properties coming from the OpenID Connect provider with the XWiki user' and provides an example of mapping, is ranked first. This directly addresses the input question about customizing user profile fields. The irrelevant nodes, discussing topics like 'access tokens and scopes', 'configuring OpenID Connect endpoints', 'group synchronization', and 'in-wiki configuration', are all ranked lower, ensuring a perfect contextual precision score. Great job on the ranking!
ContextualRecall: The score is 0.80 because most of the expected output is supported by the 3rd node in the retrieval context, which provides detailed examples and instructions on mapping user profile fields. However, the retrieval context lacks information about saving the 'xwiki.properties' file and restarting XWiki, which affects the completeness of the match.
CustomContextualRelevancy: The retrieval context includes detailed instructions on mapping user profile fields using the `oidc.user.mapping` property in the `xwiki.properties` file, matching the expected output.
Correctness: The actual output closely aligns with the expected output but includes additional details like extra placeholders (`${oidc.user.preferredUsername}`, `${oidc.user.email}`) which are not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.847586848122358
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.5
CustomContextualRelevancy: 1.0
Correctness: 0.5855210887341478
Reasons:
AnswerRelevancy: The score is 1.00 because the answer is perfectly relevant, addressing the specific question about configuring group synchronization between XWiki and an OpenID Connect provider without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions; the actual output aligns perfectly with the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because all relevant nodes are ranked higher than irrelevant nodes. The first three nodes in the retrieval context provide crucial information on configuring group synchronization, such as the `oidc.groups.claim` property and `oidc.groups.mapping`, which are essential for the process. The irrelevant nodes, ranked fourth and fifth, discuss installation instructions and authentication management, which are not directly related to the group synchronization setup. Great job on achieving perfect precision!
ContextualRecall: The score is 0.50 because while the nodes in retrieval context provide specific examples and mention properties like `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping`, they do not cover the general instructions found in the expected output.

CustomContextualRelevancy: The retrieval context contains all necessary details for configuring group synchronization, including setting the 'oidc.groups.claim', 'oidc.groups.separator', and 'oidc.groups.mapping' properties, as well as saving changes and restarting XWiki.
Correctness: The actual output includes additional information on OpenID Connect provider support and claim setup, which is not part of the expected output. It correctly mentions configuring `xwiki.properties` for group mapping and separator but uses different claim names and properties like `oidc.userinfoclaims` not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.6388471701621393
Individual_scores:
AnswerRelevancy: 0.9473684210526315
Faithfulness: 1.0
ContextualPrecision: 0.3333333333333333
ContextualRecall: 0.6
CustomContextualRelevancy: 0.3
Correctness: 0.6523812665868712
Reasons:
AnswerRelevancy: The score is 0.95 because the response is highly relevant and provides clear guidance on configuring attachment settings, but it includes an unnecessary link to basic XWiki concepts, which slightly detracts from its focus.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.33 because the relevant node in the retrieval context, ranked third, 'explains how to set the maximum size of an attachment in XWikiPreferences, which is directly relevant to configuring the maximum upload size.' However, irrelevant nodes ranked first and second 'discuss linking to external files and images, but do not address configuring upload restrictions or size limits' and 'provide a general overview of XWiki's structure and features, but do not mention attachment upload settings or size restrictions,' respectively, which should be ranked lower than the relevant node.
ContextualRecall: The score is 0.60 because while nodes in the retrieval context support specific steps for configuring mimetypes and setting the maximum upload size (as seen in nodes 3 and 4), the introductory and explanatory sentences, as well as headings, do not have direct support from any nodes.
CustomContextualRelevancy: The retrieval context mentions setting maximum attachment size and mimetype restrictions but lacks specifics on configuring 'xwiki.properties' and setting size to 10MB.
Correctness: Actual output uses the Attachment Validation Application instead of editing the xwiki.properties file for mimetype restrictions. Size configuration steps align well with expected output, and no contradictions were noted.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.6296984929444197
Individual_scores:
AnswerRelevancy: 0.6818181818181818
Faithfulness: 1.0

ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.09637277584833669
Reasons:
AnswerRelevancy: The score is 0.68 because the answer partially addresses the question of denying script rights to a space administrator, but it includes multiple irrelevant source links and a generic label that do not contribute to the solution.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant. The first node discusses programming rights and script execution but does not address the denial of script rights to a space administrator. The second node lists various rights and their properties but does not mention the inability to deny script rights to a space administrator. The third node explains the script right and its default settings but does not state that denying script rights to a space administrator is unsupported. The fourth node contains technical code snippets related to access checks but does not provide information on denying script rights to a space administrator. Finally, the fifth node discusses converting rights and excluding spaces but does not mention the denial of script rights to a space administrator.
ContextualRecall: The score is 1.00 because the expected output perfectly aligns with the information provided by the node in the retrieval context. Great job!
CustomContextualRelevancy: The expected answer is fully supported by the retrieval context, which states that rights implied by admin cannot be denied.
Correctness: The actual output contradicts the expected output by suggesting that the script right can be denied, whereas the expected output states it cannot be denied due to admin rights implications.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.7627337544134255
Individual_scores:
AnswerRelevancy: 0.8
Faithfulness: 1.0
ContextualPrecision: 0.6666666666666666
ContextualRecall: 0.875
CustomContextualRelevancy: 0.5
Correctness: 0.7347358598138862
Reasons:
AnswerRelevancy: The score is 0.80 because the response largely addresses the configuration question in XWiki, but it includes multiple irrelevant source links that do not directly contribute to solving the problem.
Faithfulness: The score is 1.00 because there are no contradictions, indicating that the actual output perfectly aligns with the retrieval context. Great job!
ContextualPrecision: The score is 0.67 because the relevant nodes in the retrieval context are mostly ranked higher than the irrelevant nodes. However, the second node, which focuses on the 'parent/child relationship and nested spaces concept', and the third node, which discusses 'string formatting and OpenID Connect properties', are ranked higher than the fourth node, which provides information on XWiki's permission system, directly relating to the configuration of access. This affects the precision score as these irrelevant nodes should be ranked lower.

ContextualRecall: The score is 0.88 because most of the expected output sentences are supported by the nodes in the retrieval context, such as nodes 4 and 6 aligning with creating groups and setting rights. However, there is a lack of specific mention in the nodes regarding adjusting rights for additional teams or spaces, which slightly affects the score.
CustomContextualRelevancy: The retrieval context mentions setting rights at different levels and inheritance of access rights, which partially aligns with the expected output's steps on setting wiki-wide and space-level rights. However, it lacks specific details about creating groups for each team and setting edit rights specifically for team spaces.
Correctness: The actual output correctly details creating groups, setting view and edit rights, and inheriting rights. However, it adds an extra step about administering rights and misses mentioning setting 'View' right specifically for 'XWikiAllGroup' and selecting 'Rights: Page & Children' option.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.6775766226412651
Individual_scores:
AnswerRelevancy: 0.6363636363636364
Faithfulness: 1.0
ContextualPrecision: 0.5
ContextualRecall: 0.7142857142857143
CustomContextualRelevancy: 0.7
Correctness: 0.51481038519824
Reasons:
AnswerRelevancy: The score is 0.64 because the response partially addresses the question about authenticating users with access tokens, but includes multiple irrelevant source links that do not contribute to answering the query.
Faithfulness: The score is 1.00 because there are no contradictions, indicating a perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.50 because relevant nodes in the retrieval context are not consistently ranked higher than irrelevant nodes. The first node, ranked 1, is irrelevant as it focuses on 'LDAP configuration' without mentioning access tokens. The second node, ranked 2, is relevant as it discusses 'configuration of authorized applications', which aligns with using access tokens for authentication. The third node, ranked 3, is irrelevant since it addresses 'OpenID Connect scopes and response types', not directly related to access tokens. The fourth node, ranked 4, is relevant because it describes 'authentication with JWT', supporting the use of access tokens. The fifth node, ranked 5, is irrelevant as it covers 'various authentication mechanisms in XWiki' without focusing on token-based authentication.
ContextualRecall: The score is 0.71 because while several key aspects like authenticating users with JWT, enabling the token authenticator, and configuring authorized applications are well-supported by nodes in the retrieval context, specific details about generating a JWT token with required claims and including the token in the 'Authorization' header are missing.
CustomContextualRelevancy: The retrieval context mentions the use of JWT for authentication and the configuration of authorized applications, but lacks details on token claims and the specific process of including the token in the request header.
Correctness: The actual output matches the expected output in terms of JWT token authentication and property settings but lacks details on claims, signing key, and 'Authorized Applications' configuration. It also incorrectly mentions a fallback authenticator, which is not in the expected output.
Question_language: en
Expected_answer_language: en

Answer_language: en

Model: AI.Models.qa_phi3_medium-128k_14b_Q4

File: qa_001_result.json

Overall_score: 0.3009989903656263
Individual_scores:
AnswerRelevancy: 0.8125
Faithfulness: 0.8
ContextualPrecision: 0
ContextualRecall: 0.0
CustomContextualRelevancy: 0.0
Correctness: 0.19349394219375757
Reasons:
AnswerRelevancy: The score is 0.81 because the response largely addresses the issue of the bell icon not being available in XWiki, but includes irrelevant source links and a statement about offline extension installation, which do not directly help resolve the specific problem.
Faithfulness: The score is 0.80 because the actual output implies that installing extensions offline might be possible through complex manual methods, whereas the retrieval context clearly states that offline installation is not supported.
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'installation instructions and prerequisites', the second node describes 'notification features and appearance', the third node lists 'dependencies', and the fourth node talks about 'customizing notification templates'. None of these address the issue of enabling notifications in the wiki, which is the focus of the input.
ContextualRecall: The score is 0.00 because the retrieval context does not provide any information related to enabling notifications in the wiki or the `xwiki.properties` file, which are crucial for the expected output.
CustomContextualRelevancy: The retrieval context lacks any information about enabling notifications through the `xwiki.properties` file, which is the core instruction in the expected output.
Correctness: The actual output provides various troubleshooting steps but lacks specific instructions on enabling notifications via `xwiki.properties`, as outlined in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.5656465255342522
Individual_scores:
AnswerRelevancy: 0.8
Faithfulness: 0.42857142857142855
ContextualPrecision: 0.75
ContextualRecall: 0.5
CustomContextualRelevancy: 0.5
Correctness: 0.4153077246340847
Reasons:
AnswerRelevancy: The score is 0.80 because the response largely addresses the issue of not receiving notifications in XWiki, but it includes irrelevant source links that do not directly contribute to solving the problem.

Faithfulness: The score is 0.43 because the actual output contains several contradictions: it incorrectly claims that nothing is watched by default, whereas earlier versions of XWiki had users watching the whole wiki by default; it states that no notifications are received until manual actions are taken, contrary to the fact that application types are enabled by default in later versions; it suggests automatic enabling of application types upon certain actions, which is not supported by the context; and it misrepresents the purpose of the 'network' tab in user profiles.
ContextualPrecision: The score is 0.75 because the first node in the retrieval context is relevant, explaining how to follow a user to receive notifications, and is ranked appropriately. However, the second and third nodes, which discuss filters, settings, and RSS feeds without addressing enabling notifications or subscribing to specific content, are ranked higher than the fourth node, which is relevant as it outlines ways to receive notifications through the Notifications Application. The fifth node, discussing the 'What's New' application, is also irrelevant and should be ranked lower.
ContextualRecall: The score is 0.50 because while the retrieval context supports details about enabling notifications through the 'alert' menu, following users, and checking the 'network' tab (as mentioned in sentences 4, 5, and 6 of the expected output), it lacks information on enabling notifications via `xwiki.properties` and subscribing to pages or users (sentences 1, 2, and 3).
CustomContextualRelevancy: The retrieval context includes details on following users and notification settings, but lacks explicit mention of enabling notifications via 'xwiki.properties' and watching pages or wikis.
Correctness: The actual output omits enabling notifications in xwiki.properties and subscribing to pages or users. It mentions watching users and application types but lacks details on enabling notifications through the alert menu. Some filtering options are covered, but the focus is not aligned with the expected output instructions.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.9004205638631171
Individual_scores:
AnswerRelevancy: 0.7272727272727273
Faithfulness: 0.8
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.8752506559059755
Reasons:
AnswerRelevancy: The score is 0.73 because the response partially addresses the question about enabling notifications for one's own actions in XWiki, but includes irrelevant source links that do not directly answer the query.
Faithfulness: The score is 0.80 because the actual output incorrectly states that the path to modify notification settings in XWiki is Administration > Notifications, which is not mentioned in the retrieval context.
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than the irrelevant ones. Great job on getting the order spot on!
ContextualRecall: The score is 1.00 because every sentence in the expected output is fully supported by the nodes in the retrieval context, demonstrating perfect alignment.
CustomContextualRelevancy: The retrieval context contains all the information found in the expected answer, including the default behavior of not receiving notifications about one's own activity unless explicitly targeted, and the ability to disable this filter in the 'Advanced filtering options' section of the notification settings.

Correctness: The Actual Output aligns with the Expected Output regarding default notification settings and user ability to change them. It provides additional clarity on how to modify these settings, which doesn't contradict the Expected Output. The mention of XWiki version 14.1+ is extra but not a factual inaccuracy.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.3925233142335871
Individual_scores:
AnswerRelevancy: 0.35294117647058826
Faithfulness: 0.7142857142857143
ContextualPrecision: 0
ContextualRecall: 0.75
CustomContextualRelevancy: 0.3
Correctness: 0.2379129946452201
Reasons:
AnswerRelevancy: The score is 0.35 because the output contains multiple irrelevant statements that do not directly address how to disable notifications in XWiki, such as providing a source link, unrelated navigation suggestions, and discussing unrelated features. However, the score is not lower because there may be some relevant information amidst the irrelevant content.
Faithfulness: The score is 0.71 because the actual output incorrectly states that all application types are enabled by default for new users in version 15.5+, while the retrieval context indicates that nothing is watched by default. Additionally, the actual output mentions an 'Advanced filtering options' section for controlling notifications about one's own activity, which is not found in the retrieval context.
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input question about disabling notifications in XWiki. For example, the first node discusses 'getting a notification RSS feed and settings for notifications', but does not mention disabling notifications or automatic page watching. Similarly, the second node talks about 'following a user and default watch settings', which is not related to disabling notifications. All nodes fail to address the specific query, leading to a score of 0.00.
ContextualRecall: The score is 0.75 because the retrieval context supports the expected output regarding automatic page watching and notification settings, as seen in nodes 2 and 5. However, it lacks explicit information about deleting existing notification filters, which affects the completeness of the match.
CustomContextualRelevancy: The retrieval context mentions default notification settings and user profile adjustments but lacks specific details about automatic page watching and deleting notification filters.
Correctness: The actual output describes notification settings and user control but lacks specifics about automatic page watching after major modifications and deleting filters from the Custom Filters list, which are key aspects of the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.8116396216730285
Individual_scores:
AnswerRelevancy: 0.8571428571428571

Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.6
CustomContextualRelevancy: 1.0
Correctness: 0.4126948728953138
Reasons:
AnswerRelevancy: The score is 0.86 because the output mostly addresses the question about configuring OpenID Connect in XWiki, but it includes an irrelevant statement about LDAP Authenticator, which is unrelated to the user's query.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which mentions the property 'oidc.enableUser=true', is ranked first. This directly addresses the configuration needed to enable user accounts on first login using OpenID Connect. The irrelevant nodes, discussing topics like domain-based authentication, general OpenID Connect information, container authentication, and LDAP configuration, are all ranked lower, ensuring precise retrieval.
ContextualRecall: The score is 0.60 because while the retrieval context accurately captures the key details about adjusting the `xwiki.properties` file and setting the `oidc.enableUser` property (as seen in nodes related to sentences 1, 2, and 3), it lacks information on saving the file, restarting the XWiki instance, and the automatic enabling of new user accounts upon first login, which are crucial for a complete understanding.
CustomContextualRelevancy: The retrieval context includes the exact property '# oidc.enableUser=true' that matches the expected answer, indicating the necessary configuration change in the xwiki.properties file.
Correctness: The actual output suggests modifying 'xwiki.cfg' instead of 'xwiki.properties' and lacks instruction to save and restart XWiki, but it correctly identifies setting 'oidc.enableUser' to true.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.9220803778494632
Individual_scores:
AnswerRelevancy: 0.9
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.8
CustomContextualRelevancy: 1.0
Correctness: 0.8324822670967791
Reasons:
AnswerRelevancy: The score is 0.90 because the answer is mostly relevant and provides useful information about customizing user profile fields in XWiki. However, it includes a source link that does not directly address the specific question about mapping additional fields, which slightly detracts from its relevance.
Faithfulness: The score is 1.00 because there are no contradictions, indicating that the actual output is perfectly aligned with the retrieval context. Great job maintaining accuracy!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which mentions 'It's possible to associate non standard properties coming from the OpenID Connect provider with the XWiki user' and provides an example of 'oidc.user.mapping=myxproperty1=${oidc.user.subject}', is ranked first. This directly relates to the input about mapping additional user profile fields. All irrelevant nodes, discussing topics like access tokens

and scopes, endpoints and HTTP methods, group synchronization, and in-wiki configuration, are ranked lower, ensuring perfect contextual precision.
ContextualRecall: The score is 0.80 because the retrieval context nodes provide detailed support for mapping user profile fields using 'oidc.user.mapping', as described in sentences 1 to 4 of the expected output. However, it lacks information about saving the file and restarting XWiki, which affects the completeness of the context.
CustomContextualRelevancy: The retrieval context contains the information about customizing user profile fields using the `oidc.user.mapping` property in the `xwiki.properties` file, matching the expected answer.
Correctness: The actual output correctly explains mapping user profile fields using oidc.user.mapping in xwiki.properties, aligning with the expected output. However, it introduces additional details about file location and administrative access that are not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.8073287990052678
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 0.8571428571428571
ContextualPrecision: 1.0
ContextualRecall: 0.4166666666666667
CustomContextualRelevancy: 1.0
Correctness: 0.5701632702220831
Reasons:
AnswerRelevancy: The score is 1.00 because the answer is perfectly relevant and directly addresses the question about configuring group synchronization between XWiki and an OpenID Connect provider. Great job!
Faithfulness: The score is 0.86 because the actual output incorrectly states that the claim to indicate the groups a user belongs to is `oidc.groups.claim`, while the retrieval context specifies it as `xwiki_groups`.
ContextualPrecision: The score is 1.00 because all relevant nodes are ranked higher than irrelevant nodes. The first three nodes in the retrieval context provide detailed information on setting properties like `oidc.groups.claim`, `oidc.groups.mapping`, and `oidc.groups.separator`, which are directly relevant to configuring group synchronization. The irrelevant nodes, ranked fourth and fifth, focus on general OpenID Connect information and authentication management, which do not address the specific group synchronization process.
ContextualRecall: The score is 0.42 because while some specific configuration instructions like setting `oidc.groups.claim` and `oidc.groups.separator` are supported by nodes in the retrieval context, many general instructions and introductory statements in the expected output are not directly found in the retrieval context.
CustomContextualRelevancy: The retrieval context contains all necessary details for configuring group synchronization, including setting the `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping` properties, as well as saving the `xwiki.properties` file, matching the expected output.
Correctness: Both outputs cover configuring 'oidc.groups.claim' and 'oidc.groups.mapping', but the actual output mentions 'oidc.userinfoclaims', which is not in the expected output. Actual output includes some key steps but lacks specific instructions about editing 'xwiki.properties' directly and restarting XWiki, which are present in expected output.
Question_language: en

Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.6530171595457864
Individual_scores:
AnswerRelevancy: 0.9285714285714286
Faithfulness: 1.0
ContextualPrecision: 0.3333333333333333
ContextualRecall: 0.7272727272727273
CustomContextualRelevancy: 0.3
Correctness: 0.6289254680972296
Reasons:
AnswerRelevancy: The score is 0.93 because the response is highly relevant to the question about limiting attachment types and sizes, but it includes a general link to XWiki basic concepts, which is not directly relevant to the specific query.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.33 because the relevant node in the retrieval context, which provides information about setting the maximum size of an attachment and mentions configuration steps in XWikiPreferences, is ranked third. However, irrelevant nodes, such as the first node discussing linking to external files and image display, and the second node explaining the structure of XWiki pages, are ranked higher than the relevant node. This misordering affects the score, as the relevant information is not prioritized.
ContextualRecall: The score is 0.73 because while many steps in the expected output are supported by nodes in the retrieval context, such as node 3 supporting the configuration steps for maximum attachment size, some introductory and specific configuration details are missing from the retrieval context.
CustomContextualRelevancy: The retrieval context mentions setting maximum attachment size and mimetype restrictions but lacks specific details on configuring 'xwiki.properties' and setting size to 10MB.
Correctness: Actual output correctly explains setting the maximum attachment size and mentions mime type configuration, but incorrectly suggests using `attachment.download.forceDownload` for mime-types instead of `attachment.upload.allowList` in `xwiki.properties`.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.6318548193996071
Individual_scores:
AnswerRelevancy: 0.8125
Faithfulness: 0.8333333333333334
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.14529558306430865
Reasons:
AnswerRelevancy: The score is 0.81 because the main content effectively addresses the question about denying script rights to a space administrator, but it is slightly lowered due to the inclusion of

multiple source links that do not directly pertain to the input question.
Faithfulness: The score is 0.83 because the actual output incorrectly claims that script right is allowed for all users at the main wiki level by default, while the retrieval context specifies that in XWiki 14.10+, script right is not given by default to all users at the main wiki level.
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses programming rights and script execution but does not address denying script rights to a space administrator. The second node lists various rights and their properties but does not mention the inability to deny script rights to a space administrator. The third node explains the script right and its default status but does not mention the inability to deny it to a space administrator. The fourth node provides code snippets related to checking access rights but does not discuss the specific issue of denying script rights to a space administrator. The fifth node is about converting rights during page migration and does not relate to denying script rights to a space administrator.
ContextualRecall: The score is 1.00 because every sentence in the expected output is fully supported by the nodes in the retrieval context, indicating perfect alignment.
CustomContextualRelevancy: The expected answer states that rights implied by admin right cannot be denied, which is supported by the retrieval context indicating that admin rights are not deniable.
Correctness: The actual output provides instructions on how to deny script rights to a space admin, contradicting the expected output that states it is not supported in XWiki. The actual output includes additional steps and information not present in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.7138053672715047
Individual_scores:
AnswerRelevancy: 0.9
Faithfulness: 0.875
ContextualPrecision: 0.6666666666666666
ContextualRecall: 0.875
CustomContextualRelevancy: 0.5
Correctness: 0.46616553696236185
Reasons:
AnswerRelevancy: The score is 0.90 because the response mostly addresses how to configure access rights in XWiki, but it includes irrelevant source links that do not directly contribute to solving the problem.
Faithfulness: The score is 0.88 because the actual output incorrectly claims that an explicit denial does not block inheritance, while the retrieval context clarifies that when a right is allowed at a given level, it is implicitly denied to others at the same level.
ContextualPrecision: The score is 0.67 because relevant nodes in the retrieval context are ranked higher than some irrelevant nodes, but not all. The first node discusses 'rights inheritance and administration features', which is relevant, and is correctly ranked higher. However, the second node, which focuses on 'the parent/child relationship and the introduction of nested pages', is irrelevant and should be ranked lower than the fourth node, which provides information on 'XWiki's permission system, including setting view and edit rights'. Additionally, the third node about 'string formatting and user identification' is also irrelevant and should be ranked lower than the sixth node, which explains 'basic rules for setting rights in XWiki'.
ContextualRecall: The score is 0.88 because most of the expected output sentences are well-supported by nodes in the retrieval context, such as creating groups, setting wiki-wide and space-level rights, and inheriting rights for child pages. However, there is a lack of specific mention in the retrieval context about adjusting rights for additional teams or spaces, which slightly affects the

score.
CustomContextualRelevancy: The retrieval context mentions rights inheritance and setting rights at different levels, which aligns with setting space-level rights and inheriting rights for child pages. However, it lacks specific details about creating groups for each team and setting the 'View' right for 'XWikiAllGroup'.
Correctness: The actual output partially aligns with the expected output by setting 'View' rights at the wiki level and restricting edit rights to respective teams, but it includes unnecessary administrator rights details and lacks specific steps for group creation and child page inheritance.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.7293358003717917
Individual_scores:
AnswerRelevancy: 0.6666666666666666
Faithfulness: 1.0
ContextualPrecision: 0.5
ContextualRecall: 0.7142857142857143
CustomContextualRelevancy: 0.8
Correctness: 0.6950624212783685
Reasons:
AnswerRelevancy: The score is 0.67 because the response partially addresses the question about authenticating users with access tokens, but includes multiple irrelevant source links and section headers that do not contribute to answering the question.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.50 because relevant nodes in the retrieval context are not consistently ranked higher than irrelevant ones. The first node, ranked 1, is irrelevant as it discusses 'LDAP configuration' and lacks mention of 'access tokens or JWT authentication.' The second node, ranked 2, is relevant because it covers 'Authorized Applications Configuration,' which pertains to using tokens for authentication. The third node, ranked 3, is irrelevant since it focuses on 'OpenID Connect scopes and response types,' not directly related to 'access tokens for authentication.' The fourth node, ranked 4, is relevant as it details 'JWT for authentication,' directly aligning with using access tokens. The fifth node, ranked 5, is irrelevant because it talks about 'different authentication mechanisms in XWiki' without specifically addressing 'access tokens or JWT.'
ContextualRecall: The score is 0.71 because while several key aspects of the expected output, such as the use of JWTs and configuration steps, are well-supported by nodes in the retrieval context, specific details like generating a JWT token with required claims and including it in the 'Authorization' header are not directly mentioned.
CustomContextualRelevancy: The retrieval context includes information about enabling the token authenticator, configuring authorized applications, and JWT token requirements, which aligns with the expected output. However, it lacks specific details about claims like expiration time, issuer, subject, and audience, and the use of the 'Authorization' header.
Correctness: The actual output correctly describes enabling the token authenticator and mentions user creation based on token information, aligning with expected output. However, it misses details on token claims, Ed25519 key signing, and configuring authorized applications.
Question_language: en
Expected_answer_language: en
Answer_language: en

Model: AI.Models.qa_phi3_mini-128k_4b_Q4

File: qa_001_result.json

Overall_score: 0.16666666666666666
Individual_scores:
AnswerRelevancy: 0.0
Faithfulness: 1
ContextualPrecision: 0
ContextualRecall: 0.0
CustomContextualRelevancy: 0.0
Correctness: 0.0
Reasons:
AnswerRelevancy: The score is 0.00 because none of the statements in the actual output address the specific issue of the 'Bell' not being available in XWiki. Instead, the output contains unrelated information, such as source links, incomplete sentences, and topics unrelated to the XWiki 'Bell' issue.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input query. The first node discusses 'installation instructions and prerequisites' without addressing the 'Bell' icon issue. The second node describes the 'functionality and appearance of the notifications bell' but lacks information on resolving its absence. The third node lists 'dependencies' unrelated to the 'Bell' icon problem, and the fourth node focuses on 'customizing notification templates' without mentioning the 'Bell' icon issue.
ContextualRecall: The score is 0.00 because there is no information in the nodes in retrieval context about enabling notifications in the wiki or the specific setting `notifications.enabled` in `xwiki.properties`.
CustomContextualRelevancy: The retrieval context does not mention enabling notifications via the `notifications.enabled` setting in `xwiki.properties`, which is the key information in the expected output.
Correctness: The actual output does not mention 'notifications.enabled' or 'xwiki.properties', and provides irrelevant content unrelated to enabling notifications.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.6648983710202968
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 0.8888888888888888
ContextualPrecision: 0.75
ContextualRecall: 0.5
CustomContextualRelevancy: 0.6
Correctness: 0.25050133723289186
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the issue of not receiving notifications in XWiki without any irrelevant information. Great job!
Faithfulness: The score is 0.89 because the actual output incorrectly claims that user interaction is required before any activity triggers a notification in XWiki 15.5+, whereas the retrieval context does not mention this requirement.
ContextualPrecision: The score is 0.75 because the first node in the retrieval context is relevant, explaining 'how to follow a user to receive notifications', and the fourth node also provides 'an overview of different notification methods in XWiki'. However, the second node, ranked higher than the fourth,

discusses 'filters and settings for notifications' without directly addressing enabling notifications, and the third node, also ranked higher than the fourth, mentions 'RSS feeds and notification settings' but lacks crucial information on enabling notifications. Additionally, the fifth node talks about the 'What's New' application, which is not directly related to the expected output.
ContextualRecall: The score is 0.50 because while there are references in the nodes in retrieval context to clicking on the 'alert' menu, following a user, and the 'network' tab in the user profile (supporting sentences 4, 5, and 6), there is no mention of enabling notifications in `xwiki.properties`, subscribing to pages or users, or watching a page, its children, or an entire wiki (missing support for sentences 1, 2, and 3).
CustomContextualRelevancy: The retrieval context includes information on following users and receiving notifications, but lacks details on enabling notifications in xwiki.properties and watching pages or wikis.
Correctness: The actual output does not mention enabling notifications in 'xwiki.properties', watching pages or users, or the 'alert' menu. It includes additional information on notification types, RSS feeds, and software updates not in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.6851851851851851
Individual_scores:
AnswerRelevancy: 0.1111111111111111
Faithfulness: 1
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.0
Reasons:
AnswerRelevancy: The score is 0.11 because the output contains numerous irrelevant statements that do not address the question about enabling notifications for one's own actions in XWiki. The presence of these unrelated details significantly detracts from the relevancy of the response.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant nodes in the retrieval context are ranked higher than the irrelevant nodes. The first node mentions that 'By default, one don’t receive notification about one’s own activity... You have the ability to disable this filter in the "Advanced filtering options" section of your own notification settings,' which directly relates to the expected output. The second node states 'By default, you won’t receive notifications for actions done by yourself. This can be changed by switching off the Own event filter,' aligning with the expected output about disabling the filter. The irrelevant nodes, such as the third node that focuses on administrators configuring notifications and does not mention user-specific settings, are ranked lower. Great job on the perfect ranking!
ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the information in the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context fully contains the expected answer's information about default notification settings and the ability to change them in the advanced filtering options of the user profile.
Correctness: The actual output contains no relevant information about notifications settings or disabling filters as mentioned in the expected output.
Question_language: en

Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.3380384813101436
Individual_scores:
AnswerRelevancy: 0.2222222222222222
Faithfulness: 1
ContextualPrecision: 0
ContextualRecall: 0.5
CustomContextualRelevancy: 0.3
Correctness: 0.006008665638639484
Reasons:
AnswerRelevancy: The score is 0.22 because the actual output contains multiple irrelevant statements that do not address the question about disabling notifications in XWiki, such as discussing unrelated topics like Python code snippets and SEO. However, the score is not zero, indicating there might be some relevant information present.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant. The first node discusses 'getting a notification RSS feed and the settings for notifications', but it does not mention disabling notifications or automatic page watching. The second node explains 'how to follow a user and default watch settings', but lacks information on disabling notifications or changing automatic page watching settings. Subsequent nodes similarly fail to address the user's query about disabling notifications or automatic page watching, focusing instead on unrelated aspects like 'filtering notifications', 'administrator level settings', and 'overriding default notification templates'.
ContextualRecall: The score is 0.50 because while the retrieval context supports the concept of automatic page watching and changing notification settings in the user profile (as seen in nodes 1 and 2 in the retrieval context), it lacks information about being notified of changes on your work and deleting existing notification filters, which are also part of the expected output.
CustomContextualRelevancy: The retrieval context mentions default watch settings and user profile settings but lacks details on automatic page watching for major modifications and deleting notification filters.
Correctness: The actual output contains irrelevant information and lacks key facts present in the expected output such as automatic page watching and notification settings.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.6019607843137255
Individual_scores:
AnswerRelevancy: 0.011764705882352941
Faithfulness: 1
ContextualPrecision: 1.0
ContextualRecall: 0.6
CustomContextualRelevancy: 1.0
Correctness: 0.0
Reasons:

AnswerRelevancy: The score is 0.01 because the output contains numerous irrelevant statements focusing on LDAP authentication and documentation, which do not address the specific OpenID Connect configuration issue in XWiki. The minimal relevance is due to the presence of some authentication-related content, but it lacks direct applicability to the question asked.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which mentions '# oidc.enableUser=true', is correctly ranked first. This node is directly relevant to enabling user accounts on first login, aligning perfectly with the input question. Great job on the ranking!
ContextualRecall: The score is 0.60 because while the node(s) in retrieval context correctly identify the property '# oidc.enableUser=true' that needs to be adjusted in the xwiki.properties file (as seen in sentences 1, 2, and 3 of the expected output), they do not mention the necessity of saving the file and restarting the XWiki instance, nor do they explicitly state that new user accounts will be automatically enabled upon first login via OpenID Connect, which are covered in sentences 4 and 5.
CustomContextualRelevancy: The retrieval context includes the exact property '# oidc.enableUser=true' mentioned in the expected output, indicating that the necessary information to enable user accounts on first login is present.
Correctness: The actual output focuses on LDAP and Azure Active Directory configurations, which do not relate to the expected output about enabling user accounts via OpenID Connect in XWiki by modifying the `xwiki.properties` file.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.6666666666666666
Individual_scores:
AnswerRelevancy: 0.0
Faithfulness: 1
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.0
Reasons:
AnswerRelevancy: The score is 0.00 because none of the statements in the actual output address the question about customizing user profile fields in XWiki. Instead, they discuss unrelated topics such as OAuth authentication in Django, SQLAlchemy models, and other irrelevant subjects.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which mentions 'It's possible to associate non standard properties coming from the OpenID Connect provider with the XWiki user', is ranked first. This directly addresses the input question about customizing user profile fields. The irrelevant nodes, discussing topics like 'access tokens and scopes', 'endpoints and HTTP methods', 'group synchronization and customization templates', and 'in-wiki configuration and multiple configurations through cookies', are all ranked lower, ensuring a perfect contextual precision score. Great job on getting it spot on!
ContextualRecall: The score is 1.00 because every sentence in the expected output is fully supported by the information in the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context includes detailed information on customizing user profile fields using the `oidc.user.mapping` property in the `xwiki.properties` file, matching the expected answer.

Correctness: The actual output does not address any elements of the expected output; it discusses OAuth and Django instead of OpenID Connect and XWiki.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.8413490470217143
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.5
CustomContextualRelevancy: 1.0
Correctness: 0.5480942821302864
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question about configuring group synchronization between XWiki and an OpenID Connect provider, with no irrelevant information included. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than the irrelevant nodes. The first three nodes provide detailed information on configuring properties like `oidc.groups.claim`, `oidc.groups.mapping`, and `oidc.groups.separator`, which are crucial for setting up group synchronization. The irrelevant nodes, ranked fourth and fifth, focus on prerequisites, installation instructions, and general authentication management, which do not directly address the group synchronization setup process. Great job on achieving perfect precision!
ContextualRecall: The score is 0.50 because while specific configuration details like 'oidc.groups.claim', 'oidc.groups.separator', and 'oidc.groups.mapping' are supported by nodes in the retrieval context, the general instructions for locating, editing, and saving the 'xwiki.properties' file are not directly found in the retrieval context.
CustomContextualRelevancy: The retrieval context contains all key information from the expected output, including configuring 'oidc.groups.claim', 'oidc.groups.separator', and 'oidc.groups.mapping' properties for group synchronization in the 'xwiki.properties' file.
Correctness: The actual output includes additional steps like enabling group synchronization, custom group synchronization, and troubleshooting that are not in the expected output. It lacks the specific details about configuring the separator and saving the file as in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.6082865190605594
Individual_scores:
AnswerRelevancy: 0.9166666666666666
Faithfulness: 0.8333333333333334
ContextualPrecision: 0.3333333333333333
ContextualRecall: 0.8181818181818182
CustomContextualRelevancy: 0.3
Correctness: 0.44820396284820496

Reasons:
AnswerRelevancy: The score is 0.92 because the answer is highly relevant to the question about configuring attachment size limits, but it includes an unnecessary link to a general guide on XWiki basic concepts, which slightly detracts from its directness.
Faithfulness: The score is 0.83 because the actual output incorrectly specifies that the 'Maximum Upload Size' field default is set in bytes for XWiki versions before 10.9RC1, while the retrieval context only mentions a default of about 32 megabytes without linking it to any specific version.
ContextualPrecision: The score is 0.33 because the relevant node in the retrieval context, which provides information about setting limits on the maximum size of an attachment and mentions the XWikiPreferences document, is ranked third. However, the first and second nodes discuss 'linking to external files and backlinks indexation' and 'the structure of XWiki pages and available actions', which are unrelated to configuring upload restrictions or size limits, and should be ranked lower than the relevant node.
ContextualRecall: The score is 0.82 because most of the expected output sentences align well with the nodes in the retrieval context, particularly regarding configuring mimetypes and setting maximum attachment size. However, the introductory sentence and specific configuration line are not directly supported by any node in the retrieval context.
CustomContextualRelevancy: The retrieval context mentions setting maximum attachment size and mimetype restrictions but lacks specific details like the 10MB limit and exact mimetype configuration steps found in the expected output.
Correctness: Actual output mentions configuring attachment size to 10MB correctly but uses `xwiki.cfg` instead of `xwiki.properties` for mimetype configuration and omits the mimetype restriction step.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.5007014621588045
Individual_scores:
AnswerRelevancy: 0.0
Faithfulness: 1
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.004208772952827291
Reasons:
AnswerRelevancy: The score is 0.00 because the output contains numerous irrelevant statements that do not address the question about denying script rights to a space administrator.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses programming rights and admin rights implications, but it does not address denying script rights to a space administrator. The second node lists various rights and their default states without mentioning the specific denial of script rights. The third node explains script rights and their default status but lacks information on denying them. The fourth node provides technical details about checking access rights but not about denying script rights. Lastly, the fifth node discusses converting rights and excluding pages or spaces, which is unrelated to denying script rights to a space administrator.
ContextualRecall: The score is 1.00 because the expected output is perfectly supported by the nodes in the retrieval context. Great job!

CustomContextualRelevancy: The retrieval context confirms that rights implied by admin cannot be denied, matching the expected output.
Correctness: The actual output is entirely unrelated to the expected output, containing irrelevant and incoherent content that does not address the topic of XWiki rights.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.6484931667977797
Individual_scores:
AnswerRelevancy: 0.8571428571428571
Faithfulness: 0.7777777777777778
ContextualPrecision: 0.6666666666666666
ContextualRecall: 1.0
CustomContextualRelevancy: 0.3
Correctness: 0.28937169919937694
Reasons:
AnswerRelevancy: The score is 0.86 because the response largely addresses the question about configuring permissions in XWiki, but includes sources that do not directly relate to the specific task of configuring permissions, which slightly detracts from its relevance.
Faithfulness: The score is 0.78 because the actual output incorrectly claims that actions can be allowed/denied at the wiki-wide level and page level by selecting 'Space', which contradicts the retrieval context that does not specify selecting 'Space' for these levels.
ContextualPrecision: The score is 0.67 because relevant nodes in the retrieval context, such as the first node discussing 'rights inheritance and administration features' and the fourth node explaining 'the permission system in XWiki', are ranked higher. However, irrelevant nodes like the second node focusing on 'parent/child relationship and the Nested Pages concept' and the third node about 'string formatting and user identifiers in OpenID Connect' are ranked higher than some relevant nodes, which prevents a higher score.
ContextualRecall: The score is 1.00 because every sentence in the expected output is well-supported by the nodes in the retrieval context. Fantastic job ensuring all details are covered!
CustomContextualRelevancy: The retrieval context mentions rights inheritance and setting rights at different levels, but lacks specific steps like creating groups for each team or setting space-level rights as detailed in the expected output.
Correctness: The actual output incorrectly suggests setting both view and edit permissions to Deny at the page level, conflicting with the expected output's guidance of allowing view access to all and restricting edit access to each team's space. It also introduces unnecessary steps not present in the expected output, like setting deny permissions for creating pages outside of assigned spaces.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.5023809523809524
Individual_scores:
AnswerRelevancy: 0.0
Faithfulness: 1
ContextualPrecision: 0.5
ContextualRecall: 0.7142857142857143

CustomContextualRelevancy: 0.8
Correctness: 0.0
Reasons:
AnswerRelevancy: The score is 0.00 because none of the statements in the actual output are relevant to the question about authenticating users with access tokens. The output contains numerous unrelated topics and lacks any direct answer to the input question.
Faithfulness: The score is 1.00 because there are no contradictions, indicating that the actual output is perfectly aligned with the retrieval context. Great job!
ContextualPrecision: The score is 0.50 because relevant nodes in the retrieval context, such as the second node which 'explains the configuration of authorized applications, which is relevant to using access tokens for authentication', and the fourth node which 'provides a description of the extension for authenticating users with JWT tokens, directly addressing the input question', are ranked lower than irrelevant nodes. The first node, discussing 'LDAP configuration and does not mention token-based authentication or JWT', and the third node, focusing on 'OpenID Connect scopes and response types, which is unrelated to the use of JWT tokens for authentication', should be ranked lower than these relevant nodes.
ContextualRecall: The score is 0.71 because while several key aspects of the expected output, such as the use of JWT for authentication (sentence 1) and configuration details (sentences 2 and 5), are supported by nodes in the retrieval context, some specific instructions like generating a JWT token with required claims (sentence 3) and including it in the 'Authorization' header (sentence 4) are not mentioned in the retrieval context.
CustomContextualRelevancy: The retrieval context confirms the use of JWT for authentication and mentions configuring authorized applications, but lacks details on token claims and specific configuration steps like setting the authclass property.
Correctness: The actual output is completely unrelated to the expected output, discussing unrelated topics like AI and Python, whereas the expected output focuses on token-based authentication for LLM Application.
Question_language: en
Expected_answer_language: en
Answer_language: en

Model: AI.Models.qa_qwen2_7b_Q4

File: qa_001_result.json

Overall_score: 0.3229596164826022
Individual_scores:
AnswerRelevancy: 0.9230769230769231
Faithfulness: 1.0
ContextualPrecision: 0
ContextualRecall: 0.0
CustomContextualRelevancy: 0.0
Correctness: 0.014680775818690081
Reasons:
AnswerRelevancy: The score is 0.92 because the answer is mostly relevant and addresses the issue of the 'Bell' not being available in XWiki. However, it includes a source link that does not directly address the specific problem, which slightly reduces the score.
Faithfulness: The score is 1.00 because there are no contradictions. Great job on maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'installation instructions and prerequisites for extensions,' which does not help with enabling notifications. The second node describes 'the notification bell and its

functionality,' but lacks information on enabling it. The third node lists 'dependencies for the notifications extension,' which is not related to the issue of enabling notifications. The fourth node explains 'how to override notification templates,' without addressing enabling notifications in the wiki settings.
ContextualRecall: The score is 0.00 because none of the sentences in the expected output can be linked to any nodes in the retrieval context, as there is no mention of enabling notifications or the specific settings in `xwiki.properties`.
CustomContextualRelevancy: The retrieval context does not mention enabling notifications or the `notifications.enabled` setting in `xwiki.properties`.
Correctness: The actual output does not address enabling notifications through the `notifications.enabled` setting in `xwiki.properties` as mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_002_result.json

Overall_score: 0.6842243413933563
Individual_scores:
AnswerRelevancy: 1.0
Faithfulness: 1.0
ContextualPrecision: 0.8055555555555555
ContextualRecall: 0.5
CustomContextualRelevancy: 0.5
Correctness: 0.2997904928045828
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating that the actual output is perfectly aligned with the retrieval context. Great job maintaining accuracy!
ContextualPrecision: The score is 0.81 because the relevant nodes in the retrieval context are mostly ranked higher than the irrelevant ones. For example, the first node is relevant as it explains how to follow a user to receive notifications, which is pertinent to the query. However, the second node, which discusses filtering notifications and hidden pages, is not directly related to enabling notifications and should be ranked lower. Similarly, the fifth node about the 'What's New' application does not address the issue of not receiving notifications and should also be ranked lower than the relevant nodes.
ContextualRecall: The score is 0.50 because while some sentences, like those about using the 'alert' menu (node 4) and following users (node 1), are supported by nodes in the retrieval context, key details such as enabling notifications in `xwiki.properties` and subscribing to pages or users are not covered.
CustomContextualRelevancy: The retrieval context includes information on following users and receiving notifications, but lacks details on enabling notifications in xwiki.properties and watching pages or wikis.
Correctness: The actual output provides troubleshooting steps but misses specific instructions like setting `notifications.enabled` to `true` in `xwiki.properties`. It also diverges into other topics like RSS feeds, which are not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_003_result.json

Overall_score: 0.9564241922299951
Individual_scores:

AnswerRelevancy: 0.8
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.9385451533799708
Reasons:
AnswerRelevancy: The score is 0.80 because the output is mostly relevant to the question about enabling notifications for one's own actions in XWiki, but it includes irrelevant source links that do not directly address the issue.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant nodes in the retrieval context are perfectly ranked higher than the irrelevant ones. The first and second nodes directly address the expected output by explaining how to enable notifications for one's own actions in XWiki. In contrast, the third node focuses on administrators configuring notifications, the fourth node discusses RSS feeds and notification types, and the fifth node lists various notification methods, none of which address the specific issue of receiving notifications for one's own actions.
ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context fully contains the expected answer, mentioning that by default one does not receive notifications about one's own activity unless explicitly targeted, and that this filter can be disabled in the 'Advanced filtering options' section of the notification settings.
Correctness: The actual output accurately describes the method to receive notifications for one's own actions and matches the expected output. Both mention accessing 'Advanced filtering options' in user settings. Minor discrepancy in wording but overall meaning aligns.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_004_result.json

Overall_score: 0.6953547654476533
Individual_scores:
AnswerRelevancy: 0.9629629629629629
Faithfulness: 0.875
ContextualPrecision: 0.5888888888888889
ContextualRecall: 1.0
CustomContextualRelevancy: 0.5
Correctness: 0.24527674083406792
Reasons:
AnswerRelevancy: The score is 0.96 because the response effectively addresses the query about disabling notifications in XWiki. However, it includes a source reference that does not directly contribute to the solution, slightly affecting the score.
Faithfulness: The score is 0.88 because the actual output incorrectly claims that XWiki allows disabling notifications about a user's own activity by default in newer versions, whereas the retrieval context states that users do not receive notifications about their own activity unless specifically targeted.
ContextualPrecision: The score is 0.59 because relevant nodes, like the second node discussing 'default watch settings' and the third node on 'filtering notifications', are ranked after an irrelevant first node that 'discusses obtaining an RSS feed for notifications'. Additionally, the fifth node, which is relevant as it 'mentions the ability for administrators to configure default notification settings', is ranked after the fourth node that 'focuses on enabling and disabling email notifications at an administrative

level', which is less relevant to the user's query.
ContextualRecall: The score is 1.00 because every sentence in the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The retrieval context mentions default watch settings and user profile settings but lacks details on automatic page watching and deleting custom filters.
Correctness: The actual output includes steps to manage notifications, which is additional information not in the expected output. It lacks the mention of automatic page watching and modifying notification filters from the 'Custom Filters' list, which are key elements in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_005_result.json

Overall_score: 0.8676662677198373
Individual_scores:
AnswerRelevancy: 0.9166666666666666
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.4
CustomContextualRelevancy: 1.0
Correctness: 0.8893309396523577
Reasons:
AnswerRelevancy: The score is 0.92 because the response is highly relevant to configuring OpenID Connect in XWiki, but it includes an irrelevant mention of LDAP authentication, which slightly detracts from the overall relevance.
Faithfulness: The score is 1.00 because there are no contradictions between the actual output and the retrieval context. Great job maintaining consistency!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which contains the property '# oidc.enableUser=true', is ranked first. This directly addresses enabling user accounts on first login, while all irrelevant nodes discussing domain-based instance authentication, session cookie configuration, prerequisites, installation instructions, container authentication, and LDAP are ranked lower.
ContextualRecall: The score is 0.40 because while the retrieval context's 1st node supports the mention of the `oidc.enableUser=true` property and its purpose, it lacks details on adding or updating properties, saving the file, restarting XWiki, and the automatic enabling of new user accounts upon first login via OpenID Connect.
CustomContextualRelevancy: The retrieval context includes the exact information needed to enable user accounts on first login by setting 'oidc.enableUser=true' in the xwiki.properties file, matching the expected output.
Correctness: The actual output accurately reflects all critical facts from the expected output, including setting the `oidc.enableUser` property to `true` and restarting XWiki. It misses the comment about uncommenting the line, which is minor.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_006_result.json

Overall_score: 0.941604774656735
Individual_scores:
AnswerRelevancy: 1.0

Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.8
CustomContextualRelevancy: 1.0
Correctness: 0.8496286479404095
Reasons:
AnswerRelevancy: The score is 1.00 because the response perfectly addresses the question without any irrelevant information. Great job!
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because the relevant node in the retrieval context, which states 'It's possible to associate non standard properties coming from the OpenID Connect provider with the XWiki user' and provides examples of 'oidc.user.mapping', is ranked first. This directly relates to mapping additional user profile fields, while the irrelevant nodes, such as the second node discussing scopes and response types, the third node about configuring endpoints and adding custom HTTP headers, the fourth node on group synchronization and customization of templates and listeners, and the fifth node on in-wiki configuration for OIDC authenticator and client configurations, are all ranked lower, maintaining perfect contextual precision. Great job on getting it spot on!
ContextualRecall: The score is 0.80 because the 3rd node in the retrieval context supports most of the expected output by detailing how to map user profile fields using the `oidc.user.mapping` property and providing examples. However, it lacks information about saving the `xwiki.properties` file and restarting XWiki, which affects the completeness of the retrieval.
CustomContextualRelevancy: The retrieval context contains all the information from the expected answer, including the ability to map additional user profile fields using the `oidc.user.mapping` property in the `xwiki.properties` file, with examples provided.
Correctness: The actual output accurately describes the process of mapping fields using `oidc.user.mapping` and provides a clear example similar to expected output. However, it introduces additional details about syntax and verification steps not mentioned in the expected output.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_007_result.json

Overall_score: 0.7915012734970306
Individual_scores:
AnswerRelevancy: 0.9130434782608695
Faithfulness: 1.0
ContextualPrecision: 1.0
ContextualRecall: 0.4
CustomContextualRelevancy: 0.9
Correctness: 0.5359641627213141
Reasons:
AnswerRelevancy: The score is 0.91 because the response is mostly relevant to configuring group synchronization between XWiki and an OpenID Connect provider. However, it includes some irrelevant details about template customization and the name of the template, which slightly detract from the focus on the synchronization process.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 1.00 because all relevant nodes in the retrieval context are ranked higher than the irrelevant nodes. The first three nodes provide detailed information on configuring the `oidc.groups.claim` property, mapping synchronized groups using `oidc.groups.mapping`, and using a

separator character in `oidc.groups.separator`, which are all crucial for synchronizing group claims between XWiki and an OpenID Connect provider. The irrelevant nodes, ranked fourth and fifth, only provide general information about OpenID Connect integration and configuration, which do not specifically address group synchronization.
ContextualRecall: The score is 0.40 because while nodes in retrieval context provide specific details about configuring properties like 'oidc.groups.claim', 'oidc.groups.separator', and 'oidc.groups.mapping', many general instructions in the expected output are not directly referenced, leading to a lower score.
CustomContextualRelevancy: The retrieval context includes details on configuring `oidc.groups.claim`, `oidc.groups.separator`, and `oidc.groups.mapping`, which align with the expected output instructions. However, it does not explicitly mention saving the `xwiki.properties` file and restarting XWiki.
Correctness: The actual output includes relevant details like configuring group claims and mapping XWiki groups, but misses specific instructions like locating and editing the `xwiki.properties` file, saving it, and restarting XWiki. It also introduces steps not present in the expected output, such as enabling debug logging and template customization.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_008_result.json

Overall_score: 0.616667065563859
Individual_scores:
AnswerRelevancy: 0.8333333333333334
Faithfulness: 1.0
ContextualPrecision: 0.3333333333333333
ContextualRecall: 0.6666666666666666
CustomContextualRelevancy: 0.3
Correctness: 0.5666690600498205
Reasons:
AnswerRelevancy: The score is 0.83 because the output provides relevant information on setting upload restrictions, but includes several irrelevant statements about forcing downloads and basic XWiki concepts, which do not directly address the question.
Faithfulness: The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!
ContextualPrecision: The score is 0.33 because the relevant node in the retrieval context, which 'provides relevant information on setting maximum attachment size and mentions the configuration of mimetype restrictions', is ranked third. Irrelevant nodes, such as the first node that 'discusses linking to external files and images, which is unrelated to setting upload restrictions on file types and sizes', should be ranked lower to improve the score.
ContextualRecall: The score is 0.67 because while several detailed steps for configuring maximum attachment size and mimetype restrictions are supported by nodes in the retrieval context, introductory and explanatory sentences, as well as step titles, are not directly found in the nodes.
CustomContextualRelevancy: The retrieval context mentions setting maximum attachment size and mimetype restrictions but lacks specifics on configuring 'xwiki.properties' for mimetypes and setting size to 10MB.
Correctness: The actual output correctly identifies setting size limits and mimetype restrictions but uses '-attachment.validation.whitelist' instead of 'attachment.upload.allowList' and lacks precise details such as '10485760' for 10MB in bytes, and mentions additional, irrelevant configurations like '-attachment.download.forceDownload'.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_009_result.json

Overall_score: 0.5600622301987723
Individual_scores:
AnswerRelevancy: 0.8333333333333334
Faithfulness: 0.5
ContextualPrecision: 0
ContextualRecall: 1.0
CustomContextualRelevancy: 1.0
Correctness: 0.027040047859300458
Reasons:
AnswerRelevancy: The score is 0.83 because the response is mostly relevant to the question about denying script rights to a space administrator. However, the inclusion of a source link that does not directly address the question slightly detracts from the overall relevancy.
Faithfulness: The score is 0.50 because the actual output incorrectly describes the conditions under which scripts can be executed in XWiki. It claims that setting the 'Script' right to 'Deny' prevents script execution unless explicitly allowed, but the retrieval context specifies a priority order for rights and conditions based on the last author's permissions. Additionally, the actual output misrepresents the default settings for script rights in the main wiki versus sub-wikis.
ContextualPrecision: The score is 0.00 because all nodes in the retrieval context are irrelevant to the input. The first node discusses 'programming rights and script execution' but not the specific issue of denying script rights to a space administrator. The second node lists 'various rights and their properties' but does not mention the inability to deny script rights to a space administrator. The third node talks about 'script rights' but not in the context of denying them to a space administrator. The fourth node is about 'checking access rights programmatically', which is unrelated. Finally, the fifth node discusses 'rights conversion during page migration', which is also unrelated.
ContextualRecall: The score is 1.00 because every part of the expected output is perfectly supported by the nodes in the retrieval context. Great job!
CustomContextualRelevancy: The expected answer states that rights implied by admin cannot be denied, which is supported by the retrieval context indicating that admin rights are not deniable and imply other rights.
Correctness: The actual output incorrectly suggests it is possible to deny the 'Script' right to space administrators, contradicting the expected output which states that rights implied by admin rights cannot be denied in XWiki.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_010_result.json

Overall_score: 0.7347222222222222
Individual_scores:
AnswerRelevancy: 0.8
Faithfulness: 0.6666666666666666
ContextualPrecision: 0.6666666666666666
ContextualRecall: 0.875
CustomContextualRelevancy: 0.5
Correctness: 0.8999999999999998
Reasons:
AnswerRelevancy: The score is 0.80 because the response largely addresses the configuration question effectively, but it includes multiple irrelevant source links that do not directly contribute to

solving the problem, preventing a higher score.
Faithfulness: The score is 0.67 because the actual output incorrectly states that a new page inherits rights from its parent space by default, whereas the retrieval context indicates that this inheritance can be overridden. Additionally, the actual output claims that wiki-wide rights take precedence over space-specific rights, which contradicts the retrieval context stating that page-level permissions override wiki-wide permissions.
ContextualPrecision: The score is 0.67 because the relevant nodes in the retrieval context, such as the first node mentioning 'rights inheritance and administration features' and the fourth node discussing XWiki's 'powerful and granular permission system', are ranked higher. However, the second node, which discusses the 'Nested Spaces concept' and naming conflicts, and the third node about 'string formatting and OpenID Connect properties', are ranked higher than the relevant fourth node, affecting the score. Additionally, the fifth node about 'creation of spaces and pages' is ranked above the relevant sixth node, which explains 'wiki wide rights, granular page level rights'.
ContextualRecall: The score is 0.88 because most of the expected output sentences are well-supported by nodes in the retrieval context, particularly regarding setting rights at different levels and managing access through groups. However, there is a lack of specific information in the retrieval context about adjusting rights for additional teams or spaces, which slightly impacts the score.
CustomContextualRelevancy: The retrieval context mentions setting rights at different levels and rights inheritance, which partially aligns with the expected output's steps on setting wiki-wide and space-level rights and inheriting rights for child pages. However, it lacks specific details about creating groups for each team and setting specific rights for XWikiAllGroup and team groups.
Correctness: The actual output closely follows the expected output with precise steps for creating groups, setting wiki-wide and space-specific rights, and inheriting rights. It adds helpful details like checking priority and testing permissions, which do not impact factual correctness.
Question_language: en
Expected_answer_language: en
Answer_language: en

File: qa_031_result.json

Overall_score: 0.7025804812124959
Individual_scores:
AnswerRelevancy: 0.7777777777777778
Faithfulness: 1.0
ContextualPrecision: 0.5
ContextualRecall: 0.7142857142857143
CustomContextualRelevancy: 0.7
Correctness: 0.523419395211484
Reasons:
AnswerRelevancy: The score is 0.78 because the response partially addresses the question about authenticating users with access tokens, but includes multiple irrelevant statements providing source links that do not directly contribute to the answer.
Faithfulness: The score is 1.00 because there are no contradictions. Great job maintaining perfect alignment with the retrieval context!
ContextualPrecision: The score is 0.50 because the relevant nodes in the retrieval context are not consistently ranked higher than irrelevant nodes. The first node, ranked 1, discusses 'LDAP authentication settings, which are unrelated to token-based authentication or JWTs,' and should be ranked lower. Similarly, the third node, ranked 3, 'discusses OpenID Connect scopes and response types, which are not directly related to the use of access tokens or JWTs for authentication,' and should also be ranked lower. However, the second node, ranked 2, 'provides information on configuring authorized applications, which is relevant to using access tokens for authentication,' and the fourth node, ranked 4, 'describes the use of JWTs for authentication, which directly relates to the expected

output about authenticating users with access tokens,' are correctly identified as relevant, contributing to the current score.
ContextualRecall: The score is 0.71 because while nodes in the retrieval context support several key aspects of the expected output, such as authenticating users with JWTs and configuring authorized applications (nodes 2 and 4), they lack information on generating JWT tokens with specific claims and including them in the 'Authorization' header.
CustomContextualRelevancy: The retrieval context contains information about enabling the token authenticator and configuring authorized applications, but lacks details on JWT claims, signing key requirements, and the specific use of the 'Authorization' header.
Correctness: The actual output correctly mentions the use of the LLM Application Authenticator and JWT for authentication, aligning with the expected output. However, it inaccurately includes OpenID Connect and omits details on token claims and Ed25519 key, which affects factual accuracy.
Question_language: en
Expected_answer_language: en
Answer_language: en

Summarization Results

Model: AI.Models.GPT4o

File: summ_001_result.json

Score: 0.4
Reason: The score is 0.40 because the summary includes extra information not present in the original text, such as the ability to print pages and the creation of multiple wikis. Additionally, the summary fails to address several questions that the original text can answer, indicating a lack of coverage and accuracy.
Score_breakdown:
Alignment: 0.6666666666666666
Coverage: 0.4
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.6
Reason: The score is 0.60 because the summary includes extra information not present in the original text, such as a tabular view summarizing rights, which may not align with the original content. Additionally, the summary fails to address specific questions that the original text can answer, indicating a lack of completeness and accuracy.
Score_breakdown:
Alignment: 0.8888888888888888
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.3333333333333333
Reason: The score is 0.33 because the summary includes a significant amount of extra information not present in the original text, such as details about managing user groups, implications of denials, and

additional resources, which suggests a lack of alignment with the original content.
Score_breakdown:
Alignment: 0.3333333333333333
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.2857142857142857
Reason: The score is 0.29 because the summary contains a significant contradiction regarding the use of the Distribution Wizard, includes multiple pieces of extra information not present in the original text, and fails to answer a question that the original text can address. These issues indicate a low-quality summarization.
Score_breakdown:
Alignment: 0.2857142857142857
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.5
Reason: The score is 0.50 because the summary contains a contradiction regarding the requirement of application installation for watchlist modifications, which is not mentioned in the original text. Additionally, it includes extra information about editing a specific page for RSS feeds that is not present in the original text. Furthermore, the summary fails to answer a question about the availability of a feed for all modified pages, which the original text can address.
Score_breakdown:
Alignment: 0.5
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.7142857142857143
Reason: The score is 0.71 because the summary includes extra information not present in the original text, such as the need for credentials from an administrator and steps for creating a page, which affects its alignment with the original content.
Score_breakdown:
Alignment: 0.7142857142857143
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.8333333333333334
Reason: The score is 0.83 because the summary includes extra information about a 'Version Control page' in XWiki that is not mentioned in the original text. However, there are no contradictions, indicating a generally accurate summary.

Score_breakdown:
Alignment: 0.8333333333333334
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.6
Reason: The score is 0.60 because the summary includes extra information not present in the original text and fails to answer specific questions that the original text can address. However, there are no contradictions, indicating some level of accuracy.
Score_breakdown:
Alignment: 0.8
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 0.8571428571428571
Reason: The score is 0.86 because the summary includes extra information about XWiki and its reference pages that is not mentioned in the original text. However, the absence of contradictions suggests a generally accurate summary.
Score_breakdown:
Alignment: 0.8571428571428571
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.8
Reason: The score is 0.80 because the summary includes extra information about examples of applications like blogs or task managers, which were not specified in the original text. However, there are no contradictions, indicating a generally accurate summary.
Score_breakdown:
Alignment: 0.8
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.GPT4o-mini

File: summ_001_result.json

Score: 0.625
Reason: The score is 0.62 because the summary includes incorrect information about the typical use of panels, adds details about wikis and sections not present in the original text, and fails to answer a question about email notifications that the original text can address.
Score_breakdown:
Alignment: 0.625

Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.6
Reason: The score is 0.60 because the summary contains a contradiction regarding where Programming Right can be granted, includes extra information about priority orders and permission precedence not mentioned in the original text, and fails to answer questions about specific rights and permissions that the original text can address.
Score_breakdown:
Alignment: 0.7272727272727273
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.45454545454545453
Reason: The score is 0.45 because the summary includes several pieces of extra information not present in the original text, such as the creation of groups for easier management and specific instructions for setting rights. Additionally, the summary fails to answer a question that the original text can address, indicating a lack of completeness and accuracy.
Score_breakdown:
Alignment: 0.45454545454545453
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.4444444444444444
Reason: The score is 0.44 because the summary contains multiple contradictions with the original text, such as incorrect methods for upgrading distribution and misrepresentations of alternative methods for starting fresh. Additionally, there is extra information in the summary not present in the original text, like troubleshooting tips, and it fails to answer questions that the original text can, such as migration paths for upgrading XWiki versions.
Score_breakdown:
Alignment: 0.4444444444444444
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.6
Reason: The score is 0.60 because the summary contains a contradiction regarding the default RSS feeds and the Watchlist Feature, and it fails to answer questions about creating new RSS feeds and receiving RSS feeds for blog posts.
Score_breakdown:
Alignment: 0.75

Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.3333333333333333
Reason: The score is 0.33 because the summary contains multiple contradictions and extra information not present in the original text, such as the scope of the Getting Started guide and details about user actions. Additionally, the summary fails to address certain questions that the original text can answer, indicating a lack of coverage and accuracy.
Score_breakdown:
Alignment: 0.3333333333333333
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.8333333333333334
Reason: The score is 0.83 because the summary includes extra information about a 'Version Control page on XWiki.org' that is not mentioned in the original text. However, there are no contradictions, indicating a generally accurate summary.
Score_breakdown:
Alignment: 0.8333333333333334
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.7142857142857143
Reason: The score is 0.71 because the summary includes extra information not present in the original text, such as accompanying images for configurations and guidance for creating a basic application. However, there are no contradictions, indicating a generally accurate representation of the original content.
Score_breakdown:
Alignment: 0.7142857142857143
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 1.0
Reason: The score is 1.00 because the summary perfectly aligns with the original text, with no contradictions or unnecessary additions.
Score_breakdown:
Alignment: 1.0
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.6666666666666666
Reason: The score is 0.67 because the summary includes extra information about specific applications like blogs, task managers, FAQs, and product management tools that are not mentioned in the original text, which slightly reduces its accuracy.
Score_breakdown:
Alignment: 0.6666666666666666
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.claude3_5_sonet

File: summ_001_result.json

Score: 0.7142857142857143
Reason: The score is 0.71 because the summary includes extra information not present in the original text, such as details about starting with a single wiki and customizing panel layouts. Additionally, the summary fails to address a question that the original text can answer regarding exporting a page in PDF format. However, there are no contradictions, indicating a generally accurate representation of the original content.
Score_breakdown:
Alignment: 0.7142857142857143
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.0
Reason: The score is 0.00 because the summary contains multiple contradictions with the original text, such as incorrectly categorizing rights and misrepresenting the permissions system. Additionally, the summary fails to answer specific questions that the original text addresses, indicating a significant lack of alignment between the summary and the original content.
Score_breakdown:
Alignment: 0.0
Coverage: 0.4
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.36363636363636365
Reason: The score is 0.36 because the summary contains significant contradictions and adds extra information not present in the original text, leading to a low-quality summarization.
Score_breakdown:
Alignment: 0.36363636363636365
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.5454545454545454
Reason: The score is 0.55 because the summary contains a significant contradiction regarding the recommended downgrading method and includes several pieces of extra information not present in the original text. These issues indicate a moderate level of accuracy and completeness in the summary.
Score_breakdown:
Alignment: 0.5454545454545454
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.75
Reason: The score is 0.75 because the summary contains a contradiction regarding the functionality of the Watchlist Feature and includes extra information about Main.WebRss and custom feeds that is not present in the original text.
Score_breakdown:
Alignment: 0.75
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.4166666666666667
Reason: The score is 0.42 because the summary includes extra information not present in the original text, such as details about login credentials, user permissions, developer access, and documentation links. Additionally, the summary fails to answer whether XWiki is described as a second generation wiki, which the original text can answer. These issues indicate a lack of alignment and completeness between the summary and the original text.
Score_breakdown:
Alignment: 0.4166666666666667
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.8888888888888888
Reason: The score is 0.89 because the summary includes extra information about XWiki.org's Version Control page that is not mentioned in the original text. However, there are no contradictions, indicating a generally accurate summary with minor additions.
Score_breakdown:
Alignment: 0.8888888888888888
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.8
Reason: The score is 0.80 because the summary includes extra information not present in the original text and fails to answer a question that the original text addresses. However, there are no contradictions, indicating a generally accurate representation of the original content.
Score_breakdown:
Alignment: 0.8571428571428571
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 0.8571428571428571
Reason: The score is 0.86 because the summary includes extra information comparing wikis to traditional content management systems, which is not mentioned in the original text. However, there are no contradictions, indicating a generally accurate summary.
Score_breakdown:
Alignment: 0.8571428571428571
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.6
Reason: The score is 0.60 because the summary includes extra information not present in the original text, such as custom applications like blogs or task managers and users customizing the platform while maintaining core functionality. However, there are no contradictions, which indicates some level of accuracy.
Score_breakdown:
Alignment: 0.6
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.command-r_35B_Q4

File: summ_001_result.json

Score: 0.4
Reason: The score is 0.40 because the summary contains contradictory information about the customization of panels, includes extra information about printing pages not mentioned in the original text, and fails to answer several questions that the original text can address, such as editing with a WYSIWYG editor, creating multiple wikis, and exporting pages in PDF format.
Score_breakdown:
Alignment: 0.6666666666666666
Coverage: 0.4
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.6
Reason: The score is 0.60 because the summary contains contradictions regarding the administration rights and editing permissions, which affects the accuracy. Additionally, the summary fails to address specific questions about default edit rights and delete rights, indicating incomplete coverage of the original text.
Score_breakdown:
Alignment: 0.6666666666666666
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.1111111111111111
Reason: The score is 0.11 because the summary contains significant contradictions and introduces numerous pieces of extra information not present in the original text. Additionally, it fails to answer questions that the original text can address, indicating a poor alignment with the original content.
Score_breakdown:
Alignment: 0.1111111111111111
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.3333333333333333
Reason: The score is 0.33 because the summary contains significant contradictions and includes extra information not found in the original text. It also fails to answer questions that the original text can address, indicating a lack of alignment and completeness.
Score_breakdown:
Alignment: 0.3333333333333333
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.6
Reason: The score is 0.60 because the summary fails to address specific questions that the original text can answer, indicating that some important details are missing from the summary. However, there are no contradictions or extra information, which suggests that the summary is generally accurate but lacks completeness.
Score_breakdown:
Alignment: 1.0
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.6

Reason: The score is 0.60 because the summary includes extra information not present in the original text, such as users learning to create pages and documentation links, and it fails to answer questions about specific features like dedicated wiki documentation and pre-installed applications, indicating a moderate level of alignment with the original text.
Score_breakdown:
Alignment: 0.6
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.4
Reason: The score is 0.40 because the summary contains significant inaccuracies, such as incorrect descriptions of features related to version comparison and reverting changes. Additionally, it introduces information not present in the original text, like a 'Version Control' page, and fails to address questions that the original text can answer, indicating a lack of coverage and precision.
Score_breakdown:
Alignment: 0.4
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.4
Reason: The score is 0.40 because the summary contains a contradiction regarding how global rights are accessed, includes extra information about screenshots not present in the original text, and fails to answer several questions that the original text can address.
Score_breakdown:
Alignment: 0.6666666666666666
Coverage: 0.4
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 0.6666666666666666
Reason: The score is 0.67 because the summary contains a contradiction regarding who can modify wiki pages and includes extra information about XWiki and references not present in the original text. Additionally, it fails to address whether most wikis include features like access rights management, which the original text can answer.
Score_breakdown:
Alignment: 0.6666666666666666
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.625

Reason: The score is 0.62 because the summary includes extra information not present in the original text, such as specific features and examples of XWiki's use cases, which were not detailed in the original text. This indicates a moderate level of alignment between the summary and the original text, but the inclusion of additional information affects the accuracy.
Score_breakdown:
Alignment: 0.625
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.gemma2_9B_Q4

File: summ_001_result.json

Score: 0.375
Reason: The score is 0.38 because the summary contains significant contradictions regarding the content of the tabbed area and panel customization, includes extra information not present in the original text such as details about XWiki as a platform and features like creating child pages, and fails to answer questions about editing and exporting capabilities that the original text covers.
Score_breakdown:
Alignment: 0.375
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.4
Reason: The score is 0.40 because the summary contains significant contradictions with the original text, such as incorrect claims about the 'createwiki' right and default permissions in XWiki. Additionally, the summary includes extra information not present in the original text, like details about register permission and inheritance rules. Furthermore, the summary fails to answer specific questions that the original text addresses, such as the default status of certain rights and the capabilities of the script right.
Score_breakdown:
Alignment: 0.7058823529411765
Coverage: 0.4
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.375
Reason: The score is 0.38 because the summary includes multiple pieces of extra information not present in the original text, such as details about permission levels, types, and behaviors. Additionally, the summary fails to address a question that the original text can answer regarding permission overrides in XWiki.
Score_breakdown:
Alignment: 0.375
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.5714285714285714
Reason: The score is 0.57 because the summary contains contradictory information about the Flavor Upgrade process and includes extra details not present in the original text, such as specific commands for upgrading and potential issues with database schema compatibility. Additionally, the summary fails to address whether XWiki provides migration paths for upgrades, which the original text can answer.
Score_breakdown:
Alignment: 0.5714285714285714
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.6666666666666666
Reason: The score is 0.67 because the summary includes extra information not found in the original text, such as the requirement of a Watchlist application for the Watchlist RSS feed and specific details about the 'Main.WebRss' page and example implementations. However, there are no contradictions, indicating a generally accurate summary with some unnecessary additions.
Score_breakdown:
Alignment: 0.6666666666666666
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.6666666666666666
Reason: The score is 0.67 because the summary includes extra information not present in the original text, such as details about creating, editing, and viewing page history, and encouragement to proceed to Step 1. Additionally, it fails to answer a question about the existence of a dedicated wiki for XWiki Rendering documentation, which the original text can answer.
Score_breakdown:
Alignment: 0.6666666666666666
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.7777777777777778
Reason: The score is 0.78 because the summary contains a minor contradiction regarding the location of the 'History' tab and includes extra information about a tutorial step not present in the original text. Despite these issues, the summary is generally accurate and informative.
Score_breakdown:
Alignment: 0.7777777777777778
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.4
Reason: The score is 0.40 because the summary includes extra information not present in the original text, such as a guide on setting user rights, screenshots illustrating configurations, and creating a basic XWiki application. Additionally, the summary fails to answer a question that the original text can, indicating a lack of coverage.
Score_breakdown:
Alignment: 0.4
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 0.7
Reason: The score is 0.70 because the summary includes extra information not present in the original text, such as the ability for anyone to edit a wiki, the concept of a constantly updated platform, and the mention of building knowledge bases and documentation. Additionally, the summary fails to address whether most wikis include features like access rights management, which the original text can answer.
Score_breakdown:
Alignment: 0.7
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.8571428571428571
Reason: The score is 0.86 because the summary accurately reflects the original text without contradictions, although it includes additional details about robust access control and user management not specified in the original text. Overall, the summary is well-aligned with the original content.
Score_breakdown:
Alignment: 0.8571428571428571
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.llama3_1_402b

File: summ_001_result.json

Score: 0.5555555555555556
Reason: The score is 0.56 because the summary contains contradictions regarding panel visibility and user control over panel width, includes extra information about the use of panels and wikis not found in the original text, and fails to answer a question about exporting pages in PDF format that the original text can address.
Score_breakdown:
Alignment: 0.5555555555555556
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.4
Reason: The score is 0.40 because the summary contains significant contradictions and adds multiple pieces of extra information not present in the original text. Additionally, it fails to answer specific questions that the original text can address, indicating a lack of fidelity and completeness in the summarization.
Score_breakdown:
Alignment: 0.5882352941176471
Coverage: 0.4
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.3333333333333333
Reason: The score is 0.33 because the summary contains a contradiction regarding the navigation method to the wiki administration page and includes numerous pieces of extra information not present in the original text, such as setting permissions for users and groups, priority order for access, and specific actions for guests, among others.
Score_breakdown:
Alignment: 0.3333333333333333
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.5714285714285714
Reason: The score is 0.57 because the summary contains multiple inaccuracies and omissions compared to the original text. It includes contradictions regarding the upgrade process and additional details not present in the original text. Furthermore, it fails to address specific questions that the original text can answer, indicating a lack of comprehensiveness and precision.
Score_breakdown:
Alignment: 0.5714285714285714
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.5
Reason: The score is 0.50 because the summary contains incorrect information about the Watchlist feature, suggesting it is a default RSS feed and used for creating customized feeds, which contradicts the original text. Additionally, the summary fails to address whether RSS feeds can be received for blog posts if the Blog application is installed, which the original text can answer.
Score_breakdown:
Alignment: 0.5
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.5
Reason: The score is 0.50 because the summary includes extra information not present in the original text, such as the division into sections for different roles, specific user actions like creating and editing a page, and a conclusion directing to the next step. Additionally, the summary fails to answer questions about logging in requirements and developers' capabilities, which the original text can address.
Score_breakdown:
Alignment: 0.5
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.8
Reason: The score is 0.80 because the summary is mostly accurate and concise, with no contradictions or unnecessary information added. However, it misses answering a specific question about whether all content added to XWiki is saved over time, indicating a slight gap in coverage.
Score_breakdown:
Alignment: 1.0
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.4
Reason: The score is 0.40 because the summary contains multiple inaccuracies and omissions. It contradicts the original text by mentioning a 'Registration' rights configuration and configurations like 'Editable/Open Wiki' that are not present in the original. Additionally, it introduces extra information about the next steps in the XWiki user guide that is not found in the original text. Furthermore, the summary fails to address a question about overriding global rights at the page level for unregistered users, which the original text can answer.
Score_breakdown:
Alignment: 0.4
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 0.6
Reason: The score is 0.60 because the summary includes extra information not present in the original text and fails to answer some questions that the original text can address. However, there are no contradictions, which is a positive aspect of the summary.
Score_breakdown:
Alignment: 0.8571428571428571
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.5
Reason: The score is 0.50 because the summary includes extra information not present in the original text, such as specific features of XWiki and its suitability for education and a wide range of use cases, which were not mentioned in the original text.
Score_breakdown:
Alignment: 0.5
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.llama3_1_8b_Q4

File: summ_001_result.json

Score: 0.3333333333333333
Reason: The score is 0.33 because the summary contains significant contradictions regarding the layout of the page areas, includes extra information not present in the original text, and fails to answer a question that the original text can address. These issues indicate a lack of accuracy and completeness in the summarization.
Score_breakdown:
Alignment: 0.3333333333333333
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.4
Reason: The score is 0.40 because the summary contains significant contradictions regarding the priority and checking order of permissions, includes vague information not present in the original text, and fails to address specific questions that the original text can answer. These issues indicate a lack of accuracy and completeness in the summarization.
Score_breakdown:
Alignment: 0.75
Coverage: 0.4
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.3333333333333333
Reason: The score is 0.33 because the summary contains significant contradictions with the original text and includes a substantial amount of extra information not present in the original text. These issues indicate a lack of accuracy and relevance in the summary compared to the original content.
Score_breakdown:
Alignment: 0.3333333333333333
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.2
Reason: The score is 0.20 because the summary contains significant contradictions and adds a substantial amount of extra information not present in the original text. Additionally, it fails to address several questions that the original text can answer, indicating a poor alignment with the source material.
Score_breakdown:
Alignment: 0.2727272727272727
Coverage: 0.2
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.875
Reason: The score is 0.88 because the summary closely aligns with the original text, with only minor discrepancies such as mentioning specific browser extensions like Firefox, which were not specified in the original text. Overall, the summary is accurate and informative.
Score_breakdown:
Alignment: 0.875
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.4
Reason: The score is 0.40 because the summary includes extra information not present in the original text, such as links to guides and resources, specific guides on creating and editing pages, and invitations to proceed with certain steps. Additionally, the summary fails to answer questions that the original text can address, such as the necessity of logging in for certain actions, whether XWiki is a second-generation wiki, and the availability of additional applications on the Extensions wiki.
Score_breakdown:
Alignment: 0.42857142857142855
Coverage: 0.4
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.75
Reason: The score is 0.75 because the summary contains a contradiction regarding the location of the 'History' menu item and includes extra information about a 'Version Control page on XWiki.org' not present in the original text. Additionally, the summary fails to address whether all content added to XWiki is saved over time, which the original text can answer. Despite these issues, the summary is still relatively accurate and informative.
Score_breakdown:
Alignment: 0.75
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.8
Reason: The score is 0.80 because the summary includes extra information not present in the original text, such as screenshots or examples of configurations, and it fails to answer a question about specific configurations that the original text can address. However, there are no contradictions, indicating a generally accurate representation of the original content.
Score_breakdown:
Alignment: 0.8571428571428571
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 0.8
Reason: The score is 0.80 because the summary is mostly accurate and concise, with no contradictions or unnecessary extra information. However, it lacks the ability to answer certain questions that the original text can, indicating room for improvement in comprehensiveness.
Score_breakdown:
Alignment: 1.0
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.6666666666666666
Reason: The score is 0.67 because the summary includes extra information about features like blogs, task managers, community hubs, and education solutions that are not mentioned in the original text. However, there are no contradictions, which is a positive aspect of the summary.
Score_breakdown:
Alignment: 0.6666666666666666
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.mistral-nemo_12b_Q4

File: summ_001_result.json

Score: 0.2
Reason: The score is 0.20 because the summary includes several inaccuracies and additional information not present in the original text. It incorrectly claims features like panel customization per space and multiple wiki creation, which are not mentioned in the original text. Furthermore, it introduces actions and functionalities that are not covered in the original text, leading to a significant deviation from the source material.
Score_breakdown:
Alignment: 0.2
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.4
Reason: The score is 0.40 because the summary contains multiple inaccuracies, including contradictions about editing rights, script execution, and administration rights, as well as extra information not present in the original text. Additionally, the summary fails to answer several questions that the original text can address, indicating a lack of coverage and accuracy.
Score_breakdown:
Alignment: 0.6
Coverage: 0.4
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.4
Reason: The score is 0.40 because the summary includes a significant amount of extra information not present in the original text, which affects its accuracy and relevance.
Score_breakdown:
Alignment: 0.4
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.14285714285714285
Reason: The score is 0.14 because the summary contains multiple contradictions with the original text, such as the incorrect use of the Distribution Wizard for upgrading flavors and the misrepresentation of recommended methods for upgrading and downgrading. Additionally, the summary includes extra information not found in the original text, like specific methods for upgrading and details about Solr issues, which further detracts from its accuracy. Furthermore, the summary fails to address several questions that the original text can answer, indicating a lack of comprehensiveness.
Score_breakdown:
Alignment: 0.14285714285714285
Coverage: 0.4
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.8888888888888888
Reason: The score is 0.89 because the summary includes extra information not present in the original text, specifically mentioning the Main.WebRss page for RSS feeds, which slightly affects the accuracy. However, there are no contradictions, indicating a generally high-quality summary.
Score_breakdown:
Alignment: 0.8888888888888888
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.6
Reason: The score is 0.60 because the summary includes incorrect information about the resources available to developers, adding the Extensions Wiki and XWiki Rendering Wiki, which are not mentioned in the original text. Additionally, the summary introduces extra information about users learning to create, edit, and view history of pages, which is not in the original text. Furthermore, the summary fails to address questions about starting a new wiki and whether XWiki is a second generation wiki, which the original text can answer.
Score_breakdown:
Alignment: 0.7142857142857143
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.6
Reason: The score is 0.60 because the summary contains a contradiction regarding how users access the 'History' menu, includes extra information about XWiki.org's Version Control page not present in the original text, and fails to answer questions about content saving and the location of the 'History' tab that the original text can address.
Score_breakdown:
Alignment: 0.7142857142857143
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.5555555555555556
Reason: The score is 0.56 because the summary contains multiple contradictions regarding the configurations described in the original text. It incorrectly combines various configurations such as Editable, Viewable, Hidden, and Protected with an Open configuration, which is not mentioned in the original text. This significantly impacts the accuracy and reliability of the summary.
Score_breakdown:
Alignment: 0.5555555555555556
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 1.0
Reason: The score is 1.00 because the summary perfectly aligns with the original text, with no contradictions or extra information present. This indicates a high-quality summarization task.
Score_breakdown:
Alignment: 1.0
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.8571428571428571
Reason: The score is 0.86 because the summary contains a minor contradiction regarding the location of built-in applications for XWiki. However, there is no extra information or unanswered questions, indicating that the summary is otherwise accurate and well-aligned with the original text.
Score_breakdown:
Alignment: 0.8571428571428571
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.mistral2_large

File: summ_001_result.json

Score: 0.5454545454545454
Reason: The score is 0.55 because the summary contains contradictions regarding the creation of subwikis and the display of panels, includes extra information not present in the original text about actions available on a Page in XWiki and starting with a single wiki, and fails to answer questions about exporting pages and choosing panels, indicating a moderate level of fidelity to the original content.
Score_breakdown:
Alignment: 0.5454545454545454
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.7
Reason: The score is 0.70 because the summary contains contradictions regarding the availability of Programming Right and Script Right compared to the original text, and it includes extra information about a table in the Security Module documentation that is not mentioned in the original text.
Score_breakdown:
Alignment: 0.7
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.5333333333333333
Reason: The score is 0.53 because the summary contains significant contradictions and includes extra information not present in the original text. The contradiction regarding who can perform actions in an open wiki and the numerous pieces of extra information indicate a lack of alignment with the original content.
Score_breakdown:
Alignment: 0.5333333333333333
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.38461538461538464
Reason: The score is 0.38 because the summary contains several contradictions with the original text, such as the purpose of the Distribution Wizard and the method for handling Solr initialization. Additionally, it includes extra information not present in the original text, like upgrading from specific packages and steps for upgrading the WAR file. There is also a question that the original text can answer but not the summary, indicating a lack of comprehensiveness.
Score_breakdown:
Alignment: 0.38461538461538464
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 1.0
Reason: The score is 1.00 because the summary perfectly aligns with the original text, with no contradictions or extraneous information present. Excellent summarization quality.
Score_breakdown:
Alignment: 1.0
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.5
Reason: The score is 0.50 because the summary includes extra information not present in the original text, such as unspecified actions requiring login and directing users to create a page, which affects the accuracy and completeness of the summary.
Score_breakdown:
Alignment: 0.5
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.625
Reason: The score is 0.62 because the summary contains contradictions regarding the location and function of the 'History' tab and menu item, includes extra details not present in the original text such as a 'Version Control page on XWiki.org' and steps about changing the logo and panels, and fails to answer a question about content saving in XWiki that the original text can address.
Score_breakdown:
Alignment: 0.625
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.5555555555555556

Reason: The score is 0.56 because the summary contains several contradictions regarding user rights and registration settings that are not aligned with the original text, and it includes extra information not present in the original text. These issues indicate a moderate level of inaccuracy and incompleteness in the summary.
Score_breakdown:
Alignment: 0.5555555555555556
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 0.8
Reason: The score is 0.80 because the summary is mostly accurate and concise, with no contradictions or unnecessary information. However, it misses answering a specific question that the original text could address, indicating a slight gap in completeness.
Score_breakdown:
Alignment: 1.0
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.8
Reason: The score is 0.80 because the summary includes extra information about the types of new applications, such as blogs or task managers, that can be created in XWiki, which is not mentioned in the original text. However, there are no contradictions, indicating a generally accurate summary.
Score_breakdown:
Alignment: 0.8
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.mixtral-8x22b

File: summ_001_result.json

Score: 0.4444444444444444
Reason: The score is 0.44 because the summary includes a significant amount of extra information not present in the original text, such as details about printing pages, creating child pages, and managing access rights. Additionally, the summary fails to address a question that the original text can answer, indicating a lack of coverage and accuracy.
Score_breakdown:
Alignment: 0.4444444444444444
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.7272727272727273

Reason: The score is 0.73 because the summary contains a contradiction regarding the similarity of 'Create Wikis Right' to programming rights, which is not mentioned in the original text. Additionally, the summary includes extra information about XWiki and a tabular view that is not present in the original text. Furthermore, the summary fails to answer a question about the default status of the comment right, which the original text can address.
Score_breakdown:
Alignment: 0.7272727272727273
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.5714285714285714
Reason: The score is 0.57 because the summary includes extra information not present in the original text, such as details about priority order for permissions, public or private configuration options, and specific documentation for setting sub-wiki access rights. Additionally, the summary fails to answer a question that the original text can, indicating a lack of completeness and accuracy.
Score_breakdown:
Alignment: 0.5714285714285714
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.5384615384615384
Reason: The score is 0.54 because the summary contains multiple contradictions and extra information not present in the original text, such as incorrect methods for upgrading and additional details about package updates and configuration backups. Additionally, the summary fails to address specific questions that the original text can answer, indicating a lack of alignment and completeness.
Score_breakdown:
Alignment: 0.5384615384615384
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.7
Reason: The score is 0.70 because the summary includes extra information not found in the original text, such as the Notifications Application, Atom, RDF, and customizing RSS content. However, there are no contradictions, which indicates a generally accurate summarization with some additional details.
Score_breakdown:
Alignment: 0.7
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.7272727272727273

Reason: The score is 0.73 because the summary includes extra information not present in the original text, such as user capabilities like creating a page, editing a page, and viewing page history, as well as the ability to create complex web applications. Additionally, it omits a detail about a link to the XWiki Development Zone. However, there are no contradictions, indicating a generally accurate representation of the original content.
Score_breakdown:
Alignment: 0.7272727272727273
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.6666666666666666
Reason: The score is 0.67 because the summary includes extra information not present in the original text, such as using the browser's back button, a 'Version Control page on XWiki.org', and a step about changing the logo and panels. Additionally, there is a minor contradiction regarding the location of the 'Compare selected versions' button.
Score_breakdown:
Alignment: 0.6666666666666666
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.75
Reason: The score is 0.75 because the summary includes extra information not present in the original text, such as leaving the Users Rights screen blank and a linked page on XWiki.org with more details and screenshots. However, there are no contradictions, indicating a generally accurate summary with some additional details.
Score_breakdown:
Alignment: 0.75
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 0.9
Reason: The score is 0.90 because the summary includes extra information not present in the original text, such as mentioning XWiki or reference pages. However, there are no contradictions, indicating a generally accurate summary with minor additions.
Score_breakdown:
Alignment: 0.9
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.8333333333333334

Reason: The score is 0.83 because the summary contains a contradiction regarding the location of built-in applications, which affects its accuracy. However, there is no extra information or unanswered questions, indicating a generally high-quality summary.
Score_breakdown:
Alignment: 0.8333333333333334
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.phi3_medium-128k_14b_Q4

File: summ_001_result.json

Score: 0.5
Reason: The score is 0.50 because the summary contains contradictions regarding the structure and default settings of XWiki, such as the starting point of wikis and the default panel arrangement. Additionally, it includes extra information about unique panels per space that is not present in the original text. Furthermore, the summary fails to address specific questions about the functionality of panels and exporting options available in XWiki, which the original text can answer.
Score_breakdown:
Alignment: 0.5
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.23076923076923078
Reason: The score is 0.23 because the summary contains significant contradictions and introduces a substantial amount of extra information not present in the original text, indicating a low level of accuracy and relevance.
Score_breakdown:
Alignment: 0.23076923076923078
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.3
Reason: The score is 0.30 because the summary contains significant contradictions with the original text, such as incorrect details about permissions and admin options. Additionally, it includes extra information not present in the original text, like the comprehensive nature of the guide and specific details on rights management. Furthermore, the summary fails to address questions that the original text can answer, such as the privileges of the wiki owner and superadmin account.
Score_breakdown:
Alignment: 0.3
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.2
Reason: The score is 0.20 because the summary contains multiple contradictions with the original text, such as incorrect methods and tips for upgrades and downgrades. Additionally, it includes extra information not present in the original text and fails to answer specific questions that the original text can address. These issues significantly reduce the accuracy and usefulness of the summary.
Score_breakdown:
Alignment: 0.2
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.5
Reason: The score is 0.50 because the summary includes a contradiction about setting up alerts on the Main.WebRss page and adds extra information about XWiki documentation on RSS feeds that is not present in the original text.
Score_breakdown:
Alignment: 0.5
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.4
Reason: The score is 0.40 because the summary includes extra information not present in the original text, such as step-by-step instructions, specific actions requiring login, and details about developer features and tools. This additional content suggests a deviation from the original text, impacting the summary's accuracy.
Score_breakdown:
Alignment: 0.4
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.8
Reason: The score is 0.80 because the summary includes extra information not present in the original text and fails to answer a question that the original text can address. However, there are no contradictions, indicating a generally accurate representation of the original content.
Score_breakdown:
Alignment: 0.8333333333333334
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.5

Reason: The score is 0.50 because the summary contains a contradiction regarding the conditions under which global rights can be overridden, includes extra information not present in the original text, and fails to address specific questions that the original text can answer.
Score_breakdown:
Alignment: 0.5
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 0.6
Reason: The score is 0.60 because the summary includes extra information not present in the original text, such as specific mentions of XWiki and its features, which were not part of the original content. Additionally, the summary fails to address questions about general wiki features like access rights management and notifications, which the original text could answer.
Score_breakdown:
Alignment: 0.6
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.5
Reason: The score is 0.50 because the summary includes contradictions regarding the location of application details and adds extra information about specific applications not mentioned in the original text.
Score_breakdown:
Alignment: 0.5
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.phi3_mini-128k_4b_Q4

File: summ_001_result.json

Score: 0.375
Reason: The score is 0.38 because the summary contains several inaccuracies and contradictions with the original text, such as incorrect placement of title and author information, misrepresentation of panel placement, and misunderstanding of what constitutes a wiki. Additionally, the summary includes extra information not present in the original text, such as adjusting column sizes and selecting specific panels. Furthermore, the summary fails to address questions that the original text can answer, such as the use of panels for lateral menus in XWiki.
Score_breakdown:
Alignment: 0.375
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.36363636363636365
Reason: The score is 0.36 because the summary contains multiple contradictions with the original text, such as incorrect default statuses for various rights and misrepresentations of access requirements. Additionally, there is extra information in the summary not present in the original text, and it fails to answer a question that the original text can address.
Score_breakdown:
Alignment: 0.36363636363636365
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.4
Reason: The score is 0.40 because the summary contains significant contradictions and includes extra information not present in the original text. It inaccurately describes the hierarchy of permissions and introduces several elements not mentioned in the original text, such as guest user permissions and sub-wiki access rights. Additionally, it fails to answer questions that the original text can address, such as the precedence of page-level permissions over wiki-wide permissions.
Score_breakdown:
Alignment: 0.4
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.0
Reason: The score is 0.00 because the summary contains multiple contradictions and extra information not present in the original text. It inaccurately describes the use of the Distribution Wizard, downgrading processes, and troubleshooting tips for Solr lock-ups. Additionally, it includes details about maintaining data integrity, specific upgrade paths, and alternative approaches that are not mentioned in the original text. Furthermore, the summary fails to address questions about migration paths and the necessity of checking release notes, which the original text can answer.
Score_breakdown:
Alignment: 0.0
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.25
Reason: The score is 0.25 because the summary includes incorrect information about the RSS notification button being provided by default, adds extra details about extensions and guides not mentioned in the original text, and fails to address a question about RSS feed features related to the Watchlist.
Score_breakdown:
Alignment: 0.25
Coverage: 0.8
Input_language: en

Summary_language: en

File: summ_006_result.json

Score: 0.42857142857142855
Reason: The score is 0.43 because the summary contains contradictions regarding the content of the Developer's Guide, includes extra instructions and details not present in the original text, and fails to answer questions related to specific documentation resources that the original text can address.
Score_breakdown:
Alignment: 0.42857142857142855
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.5
Reason: The score is 0.50 because the summary contains incorrect information about the location of the 'History' tab and adds extra details about a user guide, URL, and tutorial that are not present in the original text.
Score_breakdown:
Alignment: 0.5
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.0
Reason: The score is 0.00 because the summary contains multiple contradictions with the original text, such as incorrect claims about user rights, registration settings, and permissions. Additionally, the summary includes extra information not present in the original text, like references to an Admin Guide and detailed instructions. Furthermore, the summary fails to answer questions that the original text can, such as whether XWiki allows admins to create new user accounts.
Score_breakdown:
Alignment: 0.0
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 0.5
Reason: The score is 0.50 because the summary contains contradictions regarding how changes to pages are handled, includes extra information not present in the original text, and fails to answer specific questions that the original text can address. These issues significantly impact the accuracy and completeness of the summary.
Score_breakdown:
Alignment: 0.5
Coverage: 0.6
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.8571428571428571
Reason: The score is 0.86 because the summary includes extra information about customized functionalities like blogs or product sheet managers that are not mentioned in the original text. However, there are no contradictions, indicating a generally accurate summary.
Score_breakdown:
Alignment: 0.8571428571428571
Coverage: 1.0
Input_language: en
Summary_language: en

Model: AI.Models.qwen2_7b_Q4

File: summ_001_result.json

Score: 0.4
Reason: The score is 0.40 because the summary contains significant contradictions and extra information not found in the original text. It incorrectly describes the contents of the footer and includes additional details about XWiki that are not mentioned in the original text. Furthermore, the summary fails to answer questions that the original text can address, indicating a lack of coverage and accuracy.
Score_breakdown:
Alignment: 0.4444444444444444
Coverage: 0.4
Input_language: en
Summary_language: en

File: summ_002_result.json

Score: 0.7
Reason: The score is 0.70 because the summary contains contradictions regarding the scope and default status of certain rights, includes extra information about priority order not present in the original text, and fails to answer a question about the register right for a specific pseudo-user. These issues indicate a moderate level of accuracy and completeness in the summarization.
Score_breakdown:
Alignment: 0.7
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_003_result.json

Score: 0.14285714285714285
Reason: The score is 0.14 because the summary contains significant contradictions with the original text, such as the incorrect hierarchy of permissions and access restrictions. Additionally, the summary includes numerous pieces of extra information not found in the original text, and it fails to answer questions that the original text can address. These issues indicate a poor alignment between the summary and the original content.
Score_breakdown:
Alignment: 0.14285714285714285
Coverage: 0.8

Input_language: en
Summary_language: en

File: summ_004_result.json

Score: 0.16666666666666666
Reason: The score is 0.17 because the summary contains multiple inaccuracies and omissions. It includes contradictory information about the downgrading process, extra details not present in the original text, and fails to address several questions that the original text can answer.
Score_breakdown:
Alignment: 0.16666666666666666
Coverage: 0.2
Input_language: en
Summary_language: en

File: summ_005_result.json

Score: 0.9090909090909091
Reason: The score is 0.91 because the summary is highly accurate with no contradictions and only includes minor extra information about the flexibility of RSS feeds, which does not significantly detract from the overall quality.
Score_breakdown:
Alignment: 0.9090909090909091
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_006_result.json

Score: 0.42857142857142855
Reason: The score is 0.43 because the summary contains contradictions regarding the Developer's Guide and incorrectly suggests developers use pre-installed applications. Additionally, it includes extra information about creating and editing pages, page history, and suggestions for creating a page, which are not mentioned in the original text. Furthermore, the summary fails to answer a question about a dedicated wiki for XWiki Rendering documentation that the original text can address.
Score_breakdown:
Alignment: 0.42857142857142855
Coverage: 0.8
Input_language: en
Summary_language: en

File: summ_007_result.json

Score: 0.6666666666666666
Reason: The score is 0.67 because the summary includes some incorrect information about the location of the 'More Actions' button and adds details not found in the original text, such as the 'Version Control page on XWiki.org' and changing the logo and panels in XWiki. Additionally, the summary fails to address whether all contents added to the wiki are saved over time, which the original text can answer.
Score_breakdown:
Alignment: 0.6666666666666666
Coverage: 0.8

Input_language: en
Summary_language: en

File: summ_008_result.json

Score: 0.7777777777777778
Reason: The score is 0.78 because the summary includes extra information not present in the original text, such as references to images and a specific step in a process. However, there are no contradictions, indicating a generally accurate representation of the original content.
Score_breakdown:
Alignment: 0.7777777777777778
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_009_result.json

Score: 1.0
Reason: The score is 1.00 because the summary perfectly aligns with the original text, with no contradictions or extra information present.
Score_breakdown:
Alignment: 1.0
Coverage: 1.0
Input_language: en
Summary_language: en

File: summ_010_result.json

Score: 0.5714285714285714
Reason: The score is 0.57 because the summary contains a contradiction regarding where the list of built-in applications is detailed. Additionally, the summary includes extra information about contexts and adaptability that are not mentioned in the original text.
Score_breakdown:
Alignment: 0.5714285714285714
Coverage: 1.0
Input_language: en
Summary_language: en

Text_generation Results

Model: AI.Models.GPT4o

File: text_gen_001_result.json

Score: 0.8
Reason: The Actual Output contains all the key information accurately but is not in the JSON format as the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.9
Reason: The Actual Output is relevant and coherent, addressing the Input prompt by using all keywords in a logical sentence. It maintains fluency and quality similar to the Expected Output, with only a slight variation in the verb used, which does not impact overall quality.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.9
Reason: The Actual Output effectively simplifies the concept of photosynthesis for a younger audience, similar to the Expected Output. It maintains the key theme of converting sunlight into energy and relates it to children's experiences. The language is engaging and age-appropriate, with logical flow and grammatical correctness.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, directly addressing the features and target audience. It maintains coherence with a logical flow similar to the Expected Output, covering all key points like noise-cancelling, battery life, and comfort. The fluency is excellent, with smooth and readable text, closely matching the language quality of the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 0.9
Reason: The Actual Output is a creative story about a space adventure, directly addressing the Input. It aligns with the Expected Output by featuring a group of astronauts exploring a new planet, although it expands on the theme with additional elements like interstellar travel and alien encounters. The narrative is coherent with well-connected ideas, and the language is grammatically accurate and fluent.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 1.0
Reason: The Actual Output is a coherent and fluent poem about the beauty of nature, aligning perfectly with the Input and Expected Output by describing the wonders of the natural world through vivid imagery and themes.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 1.0
Reason: The Actual Output is a coherent and fluent news article that directly addresses the Input prompt by detailing a recent scientific discovery in quantum computing. It aligns with the Expected Output by summarizing the breakthrough effectively, maintaining logical flow and clarity throughout.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 1.0
Reason: The Actual Output is a comprehensive travel guide for Kyoto, Japan, aligning perfectly with the Input prompt's requirement to create a travel guide for a popular tourist destination. It covers top attractions, accommodations, activities, and additional travel tips, meeting the Expected Output's criteria. The content is coherent, fluent, and well-structured, effectively addressing all evaluation steps.
Prompt_language: ro
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.8
Reason: The Actual Output accurately captures the event name, dates, location, and features, aligning with the Input. However, it combines 'venue' and 'address' into one line and uses different labels compared to the Expected Output, slightly affecting coherence.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the Input for a technical audience by using precise terminology such as 'physiological processes' and 'cognitive functions.' It maintains coherence and fluency similar to the Expected Output, though it slightly lacks the depth of detail found in the Expected Output's mention of 'subjective experiences.'
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.GPT4o-mini

File: text_gen_001_result.json

Score: 0.8
Reason: The Actual Output accurately extracts the key information from the Input but does not match the Expected Output format, which is a JSON object.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.9
Reason: The Actual Output is relevant to the Input prompt, coherent, and fluent. It closely matches the Expected Output, capturing the key elements and intent, with only a minor difference in wording.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the concept of photosynthesis for a younger audience by using relatable language like 'magic trick' and 'food,' which aligns with the Input prompt. It maintains coherence and fluency with a logical flow and clear vocabulary. While it introduces the idea of energy for 'playing and having fun,' which is not in the Expected Output, it still conveys the core idea of energy conversion in a child-friendly manner.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, addressing all key features and the target audience. It maintains coherence and logical flow similar to the Expected Output, with slight variations in phrasing. The fluency is excellent, with clear and grammatically correct sentences. It meets the Expected Output in all aspects but uses different language.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 0.8
Reason: The Actual Output is a creative space adventure story, aligning with the Input prompt. It is coherent and fluent, with a clear structure and logical progression. However, it diverges from the Expected Output by focusing on a galaxy-wide quest rather than solely exploring a new planet.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 1.0
Reason: The Actual Output is a coherent and fluent poem that directly addresses the Input prompt by vividly describing the beauty of nature, aligning perfectly with the Expected Output's requirement of depicting the wonders of the natural world.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 1.0
Reason: The Actual Output is highly relevant to the Input prompt, providing a detailed news article on a recent scientific discovery. It is coherent, with a logical flow and connection of ideas, and fluent, with correct grammar and natural language. It meets the Expected Output by summarizing a groundbreaking scientific finding effectively.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, providing a comprehensive travel guide to Kyoto, a popular tourist destination. It covers attractions, transportation, cultural experiences, and tips, aligning well with the Expected Output. The content is coherent, fluent, and logically structured. However, it lacks specific mention of accommodations, which slightly deviates from the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.8
Reason: The Actual Output accurately extracts the event name, date, location, and features, addressing the Input. It aligns well with the Expected Output in terms of main ideas but uses a different format and lacks explicit separation of start and end dates.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.9
Reason: The Actual Output closely aligns with the Expected Output, effectively adapting the text for a technical audience by using terms like 'neurobiological organ' and 'cognitive processes'. It maintains coherence and fluency, with only minor differences in phrasing such as 'central substrate' instead of 'biological substrate'.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.claude3_5_sonet

File: text_gen_001_result.json

Score: 0.8
Reason: The Actual Output accurately extracts the key information from the Input and presents it clearly, maintaining relevance and coherence. However, it does not match the JSON format of the

Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.9
Reason: The Actual Output directly addresses the Input prompt and aligns well with the Expected Output in terms of relevance. It maintains coherence and logical flow, similar to the Expected Output. The language is fluent and grammatically correct, matching the naturalness of the Expected Output. Minor differences in wording do not affect the overall quality.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the concept of photosynthesis for a younger audience by using relatable language and examples, such as comparing plants to food factories and likening photosynthesis to eating lunch. It maintains relevance and coherence with the Input and Expected Output, though it adds extra analogies not present in the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, addressing key themes like noise-cancelling, battery life, and comfort for tech-savvy professionals. It is coherent and fluent, with a logical flow and good grammar. While it slightly differs in wording from the Expected Output, it meets the key elements and quality criteria effectively.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 0.7
Reason: The Actual Output is relevant to the Input prompt as it is a creative space adventure story. However, it diverges from the Expected Output by focusing on a space anomaly and first contact rather than exploring a new planet. The story is coherent and fluent, but the discrepancy in the setting affects alignment with the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 1.0

Reason: The Actual Output is a coherent and fluent poem that directly addresses the Input prompt by celebrating the beauty of nature, aligning perfectly with the Expected Output of describing the wonders of the natural world.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 0.9
Reason: The Actual Output directly addresses the Input by presenting a news article about a scientific discovery, aligning with the Expected Output of summarizing a groundbreaking finding. It is coherent, logically structured, and fluent, with clear sentence structure and grammar. The fictional note slightly detracts from full realism but does not significantly impact the article's quality.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 1.0
Reason: The Actual Output provides a comprehensive travel guide for Paris, addressing key points such as attractions, accommodations, and activities, aligning well with the Expected Output. It is detailed and coherent, with logical flow and relevant information.
Prompt_language: ro
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.8
Reason: The Actual Output addresses the Input prompt by extracting relevant event details and aligns with the Expected Output in terms of topic relevance. It is coherent and logically organized, with a slight variation in format compared to the Expected Output. The fluency is maintained with correct grammar and natural language use. However, it includes additional information not present in the Expected Output, such as 'Type' and 'Features', and doesn't separate the start and end dates.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the input for a technical audience by using precise terminology like 'central processing unit,' 'afferent sensory inputs,' and 'neurochemical signaling pathways,' aligning well with the Expected Output's focus on physiological processes and cognitive functions. It maintains coherence and fluency with complex sentence structures and advanced vocabulary.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.command-r_35B_Q4

File: text_gen_001_result.json

Score: 0.9
Reason: The Actual Output accurately extracts and presents the key information from the Input. It is relevant, coherent, and fluent. The only discrepancy is the use of colons instead of JSON format in the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.9
Reason: The Actual Output is relevant to the Input prompt, transforming the keywords into a coherent sentence. It maintains logical flow and consistency with the Expected Output, though it uses 'frolicked' instead of 'enjoyed exploring'. The sentence structure and readability are fluent and match the quality of the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.9
Reason: The Actual Output effectively simplifies the concept of photosynthesis for a younger audience, using engaging language and relatable analogies. It maintains coherence and logical flow, aligning well with the Expected Output. However, it slightly embellishes with phrases like 'superpower,' which, while engaging, diverges a bit from the straightforward nature of the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output is relevant to the Input, addressing all features like noise-cancelling, 20-hour battery life, and comfortable fit, targeted at tech-savvy professionals. It is coherent with a logical flow and consistent ideas, similar to the Expected Output. The style is fluent and natural, matching the Expected Output's readability.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 0.8
Reason: The Actual Output is relevant to the Input prompt as it tells a creative space adventure story. It is coherent and fluent, with a logical flow and good grammar. However, it focuses on a single astronaut, Luna, rather than a group as mentioned in the Expected Output, which slightly misaligns with the expected narrative.
Prompt_language: en

Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, addressing the beauty of nature as requested. It is coherent, with a logical flow and consistent ideas, aligning well with the Expected Output. Language and grammar are smooth and natural, closely matching the Expected Output's fluency.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, providing a detailed article on a recent scientific discovery. It is coherent, with a logical flow and consistent ideas, matching the Expected Output's requirement for an informative article. Minor grammatical errors like 'serendetically' slightly affect fluency.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 1.0
Reason: The Actual Output provides a comprehensive travel guide for Paris, directly addressing the Input prompt. It includes detailed sections on landmarks, restaurants, accommodations, and travel tips, aligning with the Expected Output's requirement for an informative guide. The language is fluent and coherent, maintaining a logical flow throughout the guide.
Prompt_language: ro
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.8
Reason: The Actual Output addresses the main topic and includes most key details like event name, date, and location. However, it lacks separation of start and end dates and doesn't explicitly mention 'target_audience' as in the Expected Output. The output is coherent and fluent.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.8
Reason: The Actual Output addresses the main topics of the Input, such as brain functions and interpretation of external data, aligning well with the Expected Output. It maintains coherence and logical flow, though it simplifies some technical terms like 'physiological processes' and 'biological

substrate'. The language is fluent and consistent with a technical audience.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.gemma2_9B_Q4

File: text_gen_001_result.json

Score: 0.8
Reason: The Actual Output accurately extracts the key information from the Input and aligns with the Expected Output in terms of content. However, the format differs slightly as it uses bullet points instead of JSON structure.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.9
Reason: The Actual Output maintains the main topic and context of the Input, with all keywords included. It is coherent with a logical flow similar to the Expected Output, though the order of phrases differs slightly. The fluency is high, with correct grammar and syntax.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the concept of photosynthesis for a younger audience by using relatable analogies, addressing the Input prompt. It maintains coherence and fluency, though it adds more detail than the Expected Output, which is more concise.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output aligns well with the Input prompt, covering all main features such as noise-cancelling, battery life, and comfort, targeting tech-savvy professionals. It maintains coherence with a logical flow similar to the Expected Output, though it uses a slightly different structure. The fluency is high, with clear and grammatically correct sentences, closely matching the Expected Output's readability.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 0.8

Reason: The Actual Output is a creative space adventure story, aligning well with the Input prompt. It is coherent and fluent, with a logical flow and engaging narrative. However, it focuses on an anomaly rather than directly exploring a new planet, which slightly diverges from the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 1.0
Reason: The Actual Output is highly relevant to the Input, as it is a poem about the beauty of nature. It is coherent with a logical flow and consistent ideas, aligning well with the Expected Output of describing the wonders of the natural world. The language is fluent, with excellent grammar and style, matching the quality and readability expected.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 1.0
Reason: The Actual Output is highly relevant to the Input prompt, providing a detailed news article on a recent scientific discovery about Europa's ocean. It is coherent, with logically connected ideas, and fluent, with clear language and grammar. The key elements and intent of the Expected Output are met, as it summarizes a groundbreaking finding effectively.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 1.0
Reason: The Actual Output is a comprehensive travel guide for Florence, directly addressing the Input prompt. It includes detailed sections on attractions, accommodations, and activities, aligning with the Expected Output. The content is coherent, logically structured, and fluent, matching the quality expected.
Prompt_language: ro
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.8
Reason: The Actual Output accurately extracts and presents the event details, maintaining relevance and coherence. It uses natural language and proper grammar. However, it differs slightly from the Expected Output in formatting and structure, such as combining start and end dates and including additional activities information.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the text for a technical audience by using terms like 'neural network' and 'synapses,' aligning well with the Expected Output. It maintains coherence and fluency, though it slightly diverges in phrasing compared to the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.llama3_1_402b

File: text_gen_001_result.json

Score: 0.8
Reason: The Actual Output accurately extracts the key information from the Input, matching the Expected Output in content. However, the format differs slightly as it uses a list format instead of a JSON object.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.8
Reason: The Actual Output is relevant and coherent, addressing the main topic with logical flow. It is grammatically correct and natural. However, it introduces 'chased after butterflies' instead of 'enjoyed exploring,' which slightly deviates from the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the concept of photosynthesis for a younger audience by simplifying the language and adding engaging elements, such as comparing it to a special power. It addresses the Input prompt well and maintains coherence with the Expected Output by explaining the process and its purpose. The fluency is enhanced with a friendly tone, making it suitable for ages 8-10.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, addressing all features and the target audience. It maintains coherence with a logical flow similar to the Expected Output, covering noise-cancelling, battery life, and comfort. The fluency is excellent, with proper grammar and syntax, matching the Expected Output's quality.
Prompt_language: en
Expected_answer_language: en

Generated_text_language: en

File: text_gen_005_result.json

Score: 0.8
Reason: The Actual Output is a creative space adventure story that aligns with the Input prompt. It is relevant and coherent, with a logical flow and consistent ideas. However, it diverges from the Expected Output by focusing on a cosmic exploration rather than a specific new planet, which slightly affects relevance. The language is fluent and engaging, matching the Expected Output's fluency.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 1.0
Reason: The Actual Output is a coherent and fluent poem that directly addresses the Input prompt by vividly describing the beauty of nature. It aligns well with the Expected Output by capturing the wonders of the natural world through imagery and thematic elements.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 1.0
Reason: The Actual Output is a well-structured news article about a recent scientific discovery, directly addressing the Input prompt. It provides coherent and detailed information about the discovery of Homo luzonensis, matching the Expected Output's requirement for an informative summary. The language is fluent, with clear sentence structure and grammar, aligning with the quality expected.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 1.0
Reason: The Actual Output is highly relevant to the Input prompt as it provides a detailed travel guide for Santorini, a popular tourist destination. The content is coherent, well-organized, and covers key aspects such as attractions, accommodations, and activities, aligning perfectly with the Expected Output. Additionally, the language is fluent and grammatically correct.
Prompt_language: ro
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.8
Reason: The Actual Output accurately extracts the event name, date, and location, aligning well with the Input. However, it lacks specific details such as separating the start and end dates, and omits the target audience, which are present in the Expected Output.
Prompt_language: en

Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the text for a technical audience, maintaining relevance to the Input by addressing the brain's complexity and functions. It is coherent and fluent, with logical flow and correct grammar. While it captures the key elements of the Expected Output, it slightly differs in phrasing, such as 'facilitating cognitive processes' instead of 'biological substrate for cognitive functions.'
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.llama3_1_8b_Q4

File: text_gen_001_result.json

Score: 0.8
Reason: The Actual Output contains all the key information from the Input and aligns with the main ideas. However, it does not match the Expected Output format, lacking JSON structure, which affects coherence.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.9
Reason: The Actual Output is coherent and fluent, closely aligning with the Expected Output. Both mention a sunny day, a playful cat, and a garden. The Actual Output is slightly more detailed but maintains relevance and logical flow.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the concept of photosynthesis for a younger audience by using simple language and relatable analogies, addressing the key theme of how plants use sunlight for energy. It maintains coherence and fluency, though it adds extra details about oxygen that are not in the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9

Reason: The Actual Output addresses the main topics of noise-cancelling, battery life, and comfort for tech-savvy professionals. It is coherent and fluent, with logical flow and correct grammar. The Expected Output aligns well with the Input and is consistent with the Actual Output, though the Actual Output uses slightly different phrasing.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 0.8
Reason: The Actual Output is a creative space adventure story, addressing the main topic of the Input. It includes exploration and adventure themes consistent with the Expected Output. However, it focuses on a wormhole and ancient temple rather than solely exploring a new planet. The story is coherent, grammatically correct, and fluent.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 1.0
Reason: The Actual Output is a coherent and fluent poem that directly addresses the Input prompt by vividly describing the beauty of nature, aligning well with the Expected Output's theme of the wonders of the natural world.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 1.0
Reason: The Actual Output directly addresses the topic of a recent scientific discovery, aligning with the Input request. It matches the Expected Output by providing an informative summary of a groundbreaking finding in quantum physics. The article is coherent, with a logical flow of ideas, and is grammatically correct and readable.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 1.0
Reason: The Actual Output is a comprehensive travel guide for Paris, aligning perfectly with the Input prompt. It covers top attractions, accommodations, activities, and practical tips, matching the Expected Output's requirement for an informative guide on a popular tourist destination. The content is coherent, fluent, and relevant, addressing all evaluation steps effectively.
Prompt_language: ro
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.8
Reason: The Actual Output correctly extracts the event name, dates, venue, and address, aligning with the Expected Output. However, it omits the target audience, which is a key theme in the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.8
Reason: The Actual Output effectively uses technical terms and maintains relevance to the Input prompt. It introduces terms like 'cerebral cortex' and 'afferent inputs' which align with a technical audience. However, it diverges slightly from the Expected Output by emphasizing 'spiritual experiences' instead of 'subjective experiences' and using 'central processing unit (CPU)' metaphorically, which may not align perfectly with the Expected Output's focus on biological substrates.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.mistral-nemo_12b_Q4

File: text_gen_001_result.json

Score: 0.8
Reason: The Actual Output addresses the main topic by extracting the required information, but it does not match the JSON format of the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.7
Reason: The Actual Output addresses the Input prompt and includes all keywords, aligning with the Expected Output in relevance. However, it lacks coherence and fluency compared to the Expected Output, as it does not emphasize 'playful' and changes the sentence structure. The tone and style are consistent.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the concept of photosynthesis for a younger audience by using relatable analogies and simple language. It maintains coherence and fluency throughout. However, it includes additional details like chloroplasts and oxygen release not present in the Expected Output, which focuses solely on energy conversion.
Prompt_language: en

Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output addresses the Input prompt and aligns well with the Expected Output by highlighting noise-cancelling, 20-hour battery life, and comfort for professionals. It is coherent and fluent, though it lacks the explicit mention of 'tech-savvy' and 'ergonomic fit' present in the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 0.6
Reason: The Actual Output is a creative space adventure story, aligning with the Input prompt. However, it diverges from the Expected Output as it focuses on music and a nebula rather than exploring a new planet. The story is coherent and fluent, with a logical flow and high-quality language.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 1.0
Reason: The Actual Output is a coherent and fluent poem that beautifully describes the wonders of the natural world, directly addressing the Input prompt and meeting the Expected Output's intent.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 1.0
Reason: The Actual Output is highly relevant to the Input as it provides a detailed news article about a recent scientific discovery. It is coherent, with logically connected ideas and smooth flow. The language is fluent, with correct grammar and appropriate vocabulary. The Actual Output aligns well with the Expected Output by summarizing a groundbreaking scientific finding effectively.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 1.0
Reason: The Actual Output provides a comprehensive travel guide for Paris, aligning perfectly with the Input prompt and Expected Output. It covers attractions, accommodations, activities, and travel tips, maintaining coherence and fluency throughout.
Prompt_language: en
Expected_answer_language: en

Generated_text_language: en

File: text_gen_009_result.json

Score: 0.8
Reason: The Actual Output accurately captures the event name, date, venue, and address from the Input, maintaining relevance and coherence. However, it omits the 'target_audience' detail present in the Expected Output, slightly affecting completeness.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.8
Reason: The Actual Output is relevant to the Input, addressing the complexity and functions of the brain. It maintains coherence with the Expected Output, though it uses slightly different terminology. The fluency is high, with a natural and readable style, but it lacks some of the technical specificity found in the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.mistral2_large

File: text_gen_001_result.json

Score: 0.9
Reason: The Actual Output accurately extracts and presents the key information from the Input. However, it uses a different format than the Expected Output, which affects coherence.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.9
Reason: The Actual Output is relevant to the Input prompt, maintaining coherence and fluency. It uses similar keywords and structure as the Expected Output, with minor differences in verb tense and word choice.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the concept of photosynthesis for a younger audience by using simple language and relatable analogies, maintaining relevance to the Input. It is coherent and logically structured, similar to the Expected Output, but adds more engaging elements. The fluency is high, with natural and readable language suitable for ages 8-10.
Prompt_language: en

Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, addressing all product features and the target audience. It maintains coherence with the Expected Output, providing a logical flow and consistent details. The fluency is excellent, with clear sentence structure and grammar, closely matching the Expected Output's quality.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 0.8
Reason: The Actual Output aligns with the Input by presenting a space adventure, and it introduces a group exploring new worlds, similar to the Expected Output. However, the Expected Output specifically mentions astronauts exploring a new planet, while the Actual Output involves a broader cosmic journey with a space pirate conflict. The structure and fluency are coherent and grammatically accurate, but the thematic focus slightly diverges from the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 1.0
Reason: The Actual Output is a coherent and fluent poem that directly addresses the Input prompt by vividly describing the beauty of nature, aligning well with the Expected Output of depicting the wonders of the natural world.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 1.0
Reason: The Actual Output directly addresses the Input prompt by providing a detailed news article about a recent scientific discovery, specifically a breakthrough in fusion energy. It aligns with the Expected Output by summarizing the finding in an informative manner. The article is coherent, logically structured, and maintains consistency throughout. Additionally, it is fluent, using grammatically correct and natural language, matching the Expected Output's fluency level.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 1.0

Reason: The Actual Output is a comprehensive travel guide for Bali, directly addressing the Input prompt and aligning with the Expected Output by highlighting top attractions, accommodations, and activities. It is coherent, fluent, and exceeds the quality criteria by providing detailed and relevant information.
Prompt_language: ro
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.9
Reason: The Actual Output accurately extracts all event details from the Input, maintaining relevance and coherence. It includes event name, dates, venue, address, features, and audience, aligning closely with the Expected Output. The only minor discrepancy is the format of the dates and key naming, which slightly affects fluency.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.9
Reason: The Actual Output closely follows the structure and content of the Expected Output, using technical language suitable for the audience. It maintains coherence and fluency, with only slight variations in phrasing such as 'manifestation of cognitive faculties' instead of 'serving as the biological substrate for cognitive functions.'
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.mixtral-8x22b

File: text_gen_001_result.json

Score: 0.8
Reason: The Actual Output correctly extracts and presents the key information from the Input, maintaining relevance. However, it does not match the Expected Output's JSON format, affecting coherence and fluency.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.9
Reason: The Actual Output is relevant to the Input prompt, maintaining the same topic and requirements. It is coherent with a logical flow and consistent ideas, similar to the Expected Output. The only minor difference is the use of 'enjoys' instead of 'enjoyed', which slightly affects fluency but does not significantly impact readability.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the concept of photosynthesis for a younger audience by using relatable analogies and simple language, aligning well with the Expected Output. It maintains coherence and fluency, though it introduces the idea of 'magic trick' which is not present in the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, addressing all key features and the target audience. It is coherent with a logical flow and maintains fluency with correct grammar and vocabulary. It closely aligns with the Expected Output, though it adds slightly more promotional language.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 1.0
Reason: The Actual Output is a coherent and fluent creative story about a space adventure, directly addressing the Input prompt. It follows a group of astronauts exploring a new planet, aligning well with the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 1.0
Reason: The Actual Output is a coherent and fluent poem that vividly describes the beauty of nature, aligning perfectly with the Input prompt and the Expected Output of describing the wonders of the natural world.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 1.0
Reason: The Actual Output is highly relevant to the Input prompt, aligning with the topic of a recent scientific discovery. It is coherent, with a logical flow and consistency of ideas, matching the Expected Output's requirement for an informative article. The language is fluent, grammatically correct, and natural, meeting the quality expected.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 1.0
Reason: The Actual Output directly addresses the Input by providing a comprehensive travel guide for Paris, a popular tourist destination, aligning with the Expected Output by highlighting attractions, accommodations, and activities. It is coherent, logically structured, and maintains consistency with the Input and Expected Output. The language is fluent, grammatically correct, and uses natural expressions.
Prompt_language: ro
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.8
Reason: The Actual Output accurately extracts event details such as event name, date, and location, and includes additional details about the event. However, it does not separate the start and end dates or specifically identify the target audience as in the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.9
Reason: The Actual Output effectively addresses the Input by using technical terms like 'cerebrum' and 'neural network,' aligning well with the Expected Output's focus on complexity and function. It maintains coherence and fluency with logical flow and grammatical correctness, though it slightly deviates by emphasizing 'CPU' and 'locus of cognition' instead of 'biological substrate.'
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.phi3_medium-128k_14b_Q4

File: text_gen_001_result.json

Score: 0.8
Reason: The Actual Output correctly extracts the key information from the Input and is relevant, but it does not match the JSON format of the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.9
Reason: The Actual Output is relevant to the Input prompt and coherent, with ideas logically connected. It fluently describes the scene using all keywords. The difference from the Expected Output is minor, with 'frolicking in the vibrant garden' instead of 'exploring the garden'.
Prompt_language: en

Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.8
Reason: The Actual Output effectively adapts the concept of photosynthesis for a younger audience, using engaging language and relatable analogies. It introduces the idea of plants making food with sunlight, similar to the Expected Output, but adds more detail and creativity. However, it slightly diverges from the Expected Output by including additional elements like 'solar-powered kitchen' and 'tasty vegg groves,' which, while imaginative, are not present in the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output effectively addresses the topic by targeting tech-savvy professionals and highlighting noise-cancelling, 20-hour battery life, and comfort. It is coherent, fluent, and aligns well with the Expected Output, though it includes additional details about wireless connectivity and style.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input prompt, providing a detailed and creative story about a space adventure. It is coherent with a logical flow of events and maintains fluency with grammatically correct and natural language. It exceeds the Expected Output by offering a rich narrative with character development and plot depth, while still meeting the core expectation of astronauts exploring a new planet.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 1.0
Reason: The Actual Output is a coherent and fluent poem that vividly describes the beauty and wonders of nature, aligning perfectly with the Input prompt and Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input prompt, providing a detailed article on a scientific discovery about black holes, aligning with the Expected Output. It is coherent and fluent, with a logical flow and clear language. The only minor issue is the typo in the date, '2e45', which slightly

affects the fluency.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input prompt and Expected Output, providing a comprehensive travel guide to Paris with detailed sections on attractions, accommodations, and activities. It is coherent with a logical flow and consistent structure, covering essential topics expected in a travel guide. The fluency is mostly excellent, with minor grammatical errors like 'ann cups' instead of 'annual visits.'
Prompt_language: ro
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.3
Reason: The Actual Output partially addresses the Input by extracting some event details, but includes irrelevant content about flip flops. It lacks coherence and logical flow compared to the Expected Output, which clearly structures the event details. The fluency is disrupted by the unrelated content, affecting readability.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.8
Reason: The Actual Output effectively addresses the Input by using technical language and expanding on the brain's functions, aligning well with the Expected Output. It maintains coherence and fluency but includes more detail than necessary, slightly deviating from the concise nature of the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.phi3_mini-128k_4b_Q4

File: text_gen_001_result.json

Score: 0.9
Reason: The Actual Output accurately extracts the key information from the Input and matches the Expected Output in content. However, the format differs slightly as it uses a list format instead of a JSON object.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.8
Reason: The Actual Output captures the essence of the Input and is relevant to the Expected Output. It is coherent and fluent, but it adds more detail than necessary, deviating slightly from the simplicity of the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.8
Reason: The Actual Output effectively adapts the concept of photosynthesis for a younger audience by using relatable analogies and simpler language, addressing the main topic. It is coherent and fluently explains the process, though it is more detailed than the Expected Output. The main discrepancy is the added complexity which might slightly exceed the comprehension level of the target age group.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, addressing the product features and target audience effectively. It maintains coherence with a logical flow, aligning well with the Expected Output's structure. The language is fluent, with minor verbosity compared to the Expected Output, but it remains readable and grammatically correct.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 0.9
Reason: The Actual Output is a creative story about a space adventure, aligning well with the Input prompt. It is coherent and fluent, with a logical flow and grammatical correctness. It expands on the Expected Output by detailing the exploration of a new planet, Elysium VI, by astronauts, specifically focusing on Amelia Green's journey. The story exceeds the simplicity of the Expected Output by adding depth and complexity, but remains relevant and consistent with the theme of space exploration.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 0.9
Reason: The Actual Output is a coherent and fluent poem that beautifully captures the essence of nature, aligning well with the Input prompt and Expected Output. It uses vivid imagery and metaphor to describe the wonders of the natural world, meeting the intended goals effectively.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, providing a detailed news article about a scientific discovery, which aligns with the Expected Output. It maintains coherence and logical flow, detailing the discovery, its implications, and expert opinions. The fluency is strong, with appropriate sentence structure and vocabulary. However, it is slightly verbose, which could detract from conciseness.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 0.9
Reason: The Actual Output provides a detailed travel guide for Paris, addressing the Input request effectively. It maintains logical flow and consistency, covering attractions, accommodations, and activities, which aligns with the Expected Output. The language is fluent and free of grammatical errors. However, it includes promotional content which slightly deviates from a purely informative guide.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.7
Reason: The Actual Output captures the event name, dates, and location address accurately from the Input. However, it misses separating the start and end dates, specifying the venue name, and mentioning the target audience as seen in the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.8
Reason: The Actual Output effectively adapts the Input for a technical audience by using precise terminology and detailed explanations, aligning well with the Expected Output's complexity. However, it is slightly verbose and less concise than the Expected Output, which affects coherence and fluency.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

Model: AI.Models.qwen2_7b_Q4

File: text_gen_001_result.json

Score: 0.8
Reason: The Actual Output correctly extracts the key information from the Input and aligns with the Expected Output in terms of relevance and coherence, but it lacks the JSON format required by the Expected Output.
Prompt_language: en

Expected_answer_language: en
Generated_text_language: en

File: text_gen_002_result.json

Score: 0.8
Reason: The Actual Output is relevant to the Input, addressing the main topic of a cat in a garden on a sunny day. It is coherent, but lacks the adjective 'playful' from the Expected Output, affecting the alignment. The fluency is good, matching the Expected Output's quality.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_003_result.json

Score: 0.9
Reason: The Actual Output effectively adapts the concept of photosynthesis for a younger audience by using simple language and engaging imagery, aligning well with the Input prompt. It maintains coherence and fluency, with logical connections and correct grammar. While it introduces creative elements like 'secret recipe' and 'superpower,' it still conveys the main idea of photosynthesis, similar to the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_004_result.json

Score: 0.9
Reason: The Actual Output directly addresses the Input prompt by describing the wireless Bluetooth headphones with noise-cancelling, 20-hour battery life, and comfortable fit, targeting tech-savvy professionals. It maintains coherence and fluency with a logical flow and smooth language. While it includes additional details like touch controls and a charging case not mentioned in the Expected Output, it still meets the key elements and intent of the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_005_result.json

Score: 0.9
Reason: The Actual Output is a creative story about a space adventure, aligning well with the Input prompt. It is coherent, with a logical flow from the mission's start to the discovery of ancient beings. The language is fluent and grammatically correct. Although it exceeds the Expected Output by including more details about the aliens and diplomacy, it still maintains relevance to the theme of exploring a new planet.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_006_result.json

Score: 1.0

Reason: The Actual Output is a coherent and fluent poem that directly addresses the Input prompt by beautifully describing the wonders of the natural world, aligning perfectly with the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_007_result.json

Score: 1.0
Reason: The Actual Output is highly relevant to the Input, providing a detailed news article about a recent scientific discovery in neuroscience. It is coherent, with logically connected ideas and smooth, natural language. The key elements and intent of the Expected Output are met, as the article is informative and summarizes a groundbreaking scientific finding effectively.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_008_result.json

Score: 0.9
Reason: The Actual Output provides a comprehensive travel guide for Bali, addressing the Input prompt effectively with detailed sections on transportation, accommodation, attractions, and more. It maintains a logical flow and consistency in ideas, aligning well with the Expected Output of highlighting top attractions and activities. The fluency is high with clear sentence structure and grammar, matching the quality expected.
Prompt_language: ro
Expected_answer_language: en
Generated_text_language: en

File: text_gen_009_result.json

Score: 0.8
Reason: The Actual Output accurately extracts event details such as name, dates, location, and activities, aligning well with the Input. However, it misses the specific start and end dates and the target audience, which are present in the Expected Output.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en

File: text_gen_010_result.json

Score: 0.9
Reason: The Actual Output is highly relevant to the Input, addressing the complexity and functions of the brain with technical details like neuron count. It maintains coherence with the Expected Output by discussing cognitive functions and consciousness, though it introduces additional philosophical elements. The fluency is consistent with the Expected Output, using sophisticated language appropriate for a technical audience.
Prompt_language: en
Expected_answer_language: en
Generated_text_language: en