



Code

Issues

Pull requests

Projects

Wiki

Security

Insights



workshop-telkom-partnership-feb2025

/ Lab 6 (Optional) - Evaluate prompt using watsonx.governance /



lourensius-bisma Update Lab 2 - Evaluate your GenAI Model using Robust Metrics.md

94c2b9d · 4 days ago

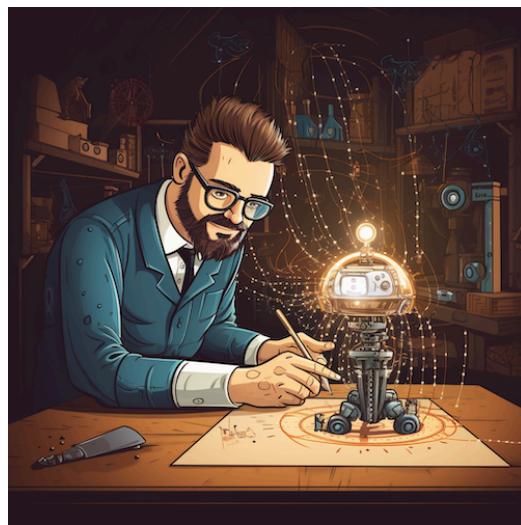


Name	Name	Last commit date
..		
data	modify	4 days ago
images	modify	4 days ago
.DS_Store	modify	4 days ago
Lab 2 - Evaluate your ...	Update Lab 2 - Evaluate you...	4 days ago
readme.md	Update readme.md	4 days ago

readme.md



Lab 6: Evaluate your GenAI Model using Robust Metrics



Pendahuluan

Di lab ini, Anda akan berperan sebagai **Prompt Engineer** dan menggunakan watsonx.governance untuk membuat *prompt template*, mengevaluasinya terhadap kumpulan data, dan melihat metrik kinerja. Setelah itu, Anda akan membuka *AI factsheet* dari *prompt* Anda untuk mempelajari cara mengubah ambang batas evaluasi, lalu bereksperimen dengan berbagai varian *prompt template* Anda dan mencoba mencapai evaluasi terbaik yang dapat Anda lakukan pada kumpulan data kami.

Table of Contents

1. [Membuat Prompt Template](#)
 - i. [Prompt Variables](#)
 - ii. [Saving Prompt](#)
 - iii. [Model Parameters](#)
2. [Mengevaluasi Prompt](#)
3. [Cara Alternatif untuk Melakukan Evaluasi Prompt](#)
4. [Hasil Evaluasi](#)
5. [Percobaan dengan Prompt Baru](#)
 - i. [Memakai Granite](#)
 - ii. [Lainnya!](#)
6. [Percobaan dengan Seting](#)

1. Membuat Prompt Template

Mari kita mulai membuat prompt pertama kita. Buka tab "Assets" di beranda proyek. Setelah itu, klik "New Asset".

IBM watsonx

Projects / Insurance Claim Summarization ...

Overview Assets Jobs Manage

Find assets Import assets New asset +

0 asset All assets

All assets

Asset types

After you create assets, they are organized by asset type.

Start working

To get started with assets, click **New asset** to create one in a tool, or **Import assets** to add existing ones.

Data in this project

Drop data files here or browse for files to upload

Pada menu yang muncul, cari kotak bertuliskan "Chat and build prompts with foundation models" untuk membuka *Prompt Lab*.

What do you want to do?

Select a task based on your goal. You'll use a tool to create an asset for that goal.

All

Prepare data

Work with models

Automate model lifecycles

Search for a task or tool

Recents

[..]

Chat and build prompts with foundation models
with Prompt Lab

[..]

Work with data and models in Python or R notebooks
with Jupyter notebook editor

Prepare data

Connect to a data source

Ground gen AI with vectorized documents

Prepare and visualize data

Define reusable sets of parameters

Setelah Anda memilih kotak tersebut, Anda akan disambut oleh layar seperti pada gambar di bawah. Halaman tersebut memiliki banyak pilihan untuk mengatur *prompt* sesuai kebutuhan Anda. Anda akan melihat bahwa ada kotak berlabel "Instruction" yang merupakan tempat kita akan meletakkan sebagian besar prompt untuk dievaluasi nanti. Tepat di atasnya, Anda melihat opsi untuk memilih mode "Structured" atau "Freeform". Lab ini akan membahas penggunaan fungsi prompt terstruktur, namun Anda juga dapat melakukannya pada mode *freeform*, jadi pilih opsi mana pun yang paling nyaman bagi Anda! Prompt terstruktur memiliki bagian yang mudah diidentifikasi untuk meletakkan contoh, instruksi, dan variabel *prompt* (yang akan kita bahas nanti). Prompt bentuk bebas memiliki struktur yang lebih longgar dan lebih fleksibel.

Anda juga akan melihat ikon dengan grafik di kanan atas, yang merupakan tombol untuk menjalankan evaluasi yang akan kita bahas nanti. Anda juga dapat melihat status pekerjaan Anda di kanan atas, yang menunjukkan apakah pekerjaan sesi prompt Anda Tersimpan atau Tidak Tersimpan. Perhatikan bahwa untuk menjalankan evaluasi, Anda perlu mengeklik menu dropdown dan menyimpan pekerjaan Anda.

The screenshot shows the IBM Watsonx interface for creating prompts. In the 'Set up' section, there is an 'Instruction (optional)' field containing the text: 'Tell the model what to do. For example: Summarize the transcript.' Below it, there is an 'Examples (optional)' section with two rows for input and output. The first row has 'Input: Enter your example input here.' and 'Output: Enter your desired output.'. A button 'Add example +' is located below this row. At the bottom left, there is a 'Try ^' button.

Di dalam kotak Instruksi, masukkan *prompt* pertama yang akan kita evaluasi di sepanjang lab ini. Untuk saat ini, masukkan perintah yang ditampilkan di layar di bawah ini.

You are an insurance agent tasked to assess insurance claims. Summarize the following insurance claim input. Focus on the car and the damage. Make the summary at least 3 sentences long.

The screenshot shows the 'Set up' section with the same instruction and examples as before. On the right side, there is a 'Model parameters' sidebar. It includes sections for 'Decoding' (set to 'Greedy'), 'Repetition penalty' (set to 1), and 'Stopping criteria'. Under 'Stopping criteria', there is a 'Stop sequences' field with a plus sign and a 'Min tokens' field set to 1. There is also a 'Max tokens' field set to 20.

Sekarang, kita perlu memastikan bahwa kita dapat menguji *prompt* kita terhadap data validasi dan data pengujian. Untuk melakukannya, kita perlu memastikan bahwa kita telah menyiapkan variabel perintah untuk instruksi kita. Dengan menggulir ke bawah halaman, Anda akan melihat area di layar berjudul "Try". Di dalamnya, Anda harus meletakkan variabel di dalam kotak Input dan memberinya nama "{input}", kemudian klik tombol "Prompt variables".

1.1 Prompt Variables

The screenshot shows the IBM Watsonx Prompt Lab interface. In the 'Set up' section, there is a 'Hint' message: 'Hint: This model works better when you provide at least 1 example.' Below it, the 'Instruction (optional)' field contains: 'You are an insurance agent tasked to assess insurance claims. Summarize the following insurance claim input. Focus on the car and the damage. Make the summary at least 3 sentences long.' The 'Examples (optional)' section has an 'Input' field with placeholder 'Enter your example input here.' and an 'Output' field with placeholder 'Enter your desired output.' A blue button 'Add example +' is visible. In the 'Try' section, there is a 'Test your prompt' field containing '{input}' and an 'Output' field showing 'Generated output appears here.' A blue 'Generate' button is at the bottom right.

Setelah bilah sisi terbuka, Anda akan melihat dua kotak kosong, berlabelkan "Variable" dan "Default Value", yang muncul dan dapat diisi. Masukkan kata "input" ke dalam kotak variabel, dan "null" ke dalam kotak "Default Value".

The screenshot shows the IBM Watsonx Prompt Lab interface with additional configuration on the right side. A sidebar titled 'Prompt variables' lists a single entry: 'Variable' 'input' and 'Default value' 'null'. A blue 'Preview' button is also present in this sidebar. The main interface remains the same as the previous screenshot, showing the 'Set up' and 'Try' sections.

Di samping tombol *prompt variables*, Anda akan menemukan tombol yang bertuliskan "Model parameters". Kita tidak akan melakukan banyak perubahan untuk saat ini, hanya perkenalan sehingga Anda bisa melakukan modifikasi nanti. Hal terpenting saat ini adalah pada saat pertama kali membuka tab, Anda akan melihat tombol "Greedy" dan "Sampling". Memilih *Greedy* akan memungkinkan hasil model Anda dapat direproduksi di seluruh eksperimen, sementara memilih *Sampling* akan memberi Anda keluaran yang lebih tinggi tingkat variabilitasnya. Untuk tugas membuat ringkasan saat ini, kita akan memilih opsi Sampling untuk *prompt* dasar kita.

1.2 Model Parameters

The screenshot shows the IBM WatsonX Prompt Lab interface. On the right side, under the 'Model parameters' section, the 'Sampling' option is selected, indicated by a blue toggle switch. Other options like 'Greedy' and 'Decoding' are shown with greyed-out toggles. Below the 'Sampling' section are sliders for 'Repetition penalty' (set to 1), 'Temperature' (set to 1.0), and 'Top P (nucleus sampling)' (set to 1). There are also fields for 'Min tokens' (0) and 'Max tokens' (20). A note at the bottom states: 'Enter up to 6 sequences to stop output after the minimum number of tokens is reached.'

Setelah Anda memilih *Sampling*, Anda akan melihat beberapa opsi baru yang muncul, khusus untuk metode Sampling. Suhu (temperature) merupakan faktor terpenting di sini, di mana pengaturan suhu yang lebih tinggi akan memberikan variasi yang lebih tinggi dan keluaran yang lebih unik. Coba atur parameter ini ke nilai 1 seperti yang saya lakukan pada gambar berikut. Lalu, arahkan kursor ke ikon info untuk mendapatkan informasi lebih lanjut tentang parameter Top P dan Top K.

This screenshot shows the same IBM WatsonX Prompt Lab interface as above, but with specific parameter values highlighted. The 'Temperature' slider is set to 1.0. The 'Top P (nucleus sampling)' slider is set to 1. The 'Top K' slider is set to 50. The 'Random seed' field is empty. The rest of the interface remains the same, including the 'Instruction (optional)' text area and the 'Examples (optional)' section.

Gulir ke bawah untuk menemukan parameter *Stopping criteria* dan *Max tokens*. Nilai awal untuk token maksimum adalah 20. Untuk LLM, token tidak selalu sama dengan kata, jadi kita ingin menetapkannya lebih tinggi (misalnya 200) untuk mendapatkan lebih banyak detail pada ringkasan klaim asuransi kita.

The screenshot shows the WatsonX Prompt Lab interface. In the center, there's a 'Test your prompt' section with an 'Input' field containing '{input}' and an 'Output' field showing 'Generated output appears here.' Below this is a 'New test' button. To the right, there are several configuration sliders and input fields:

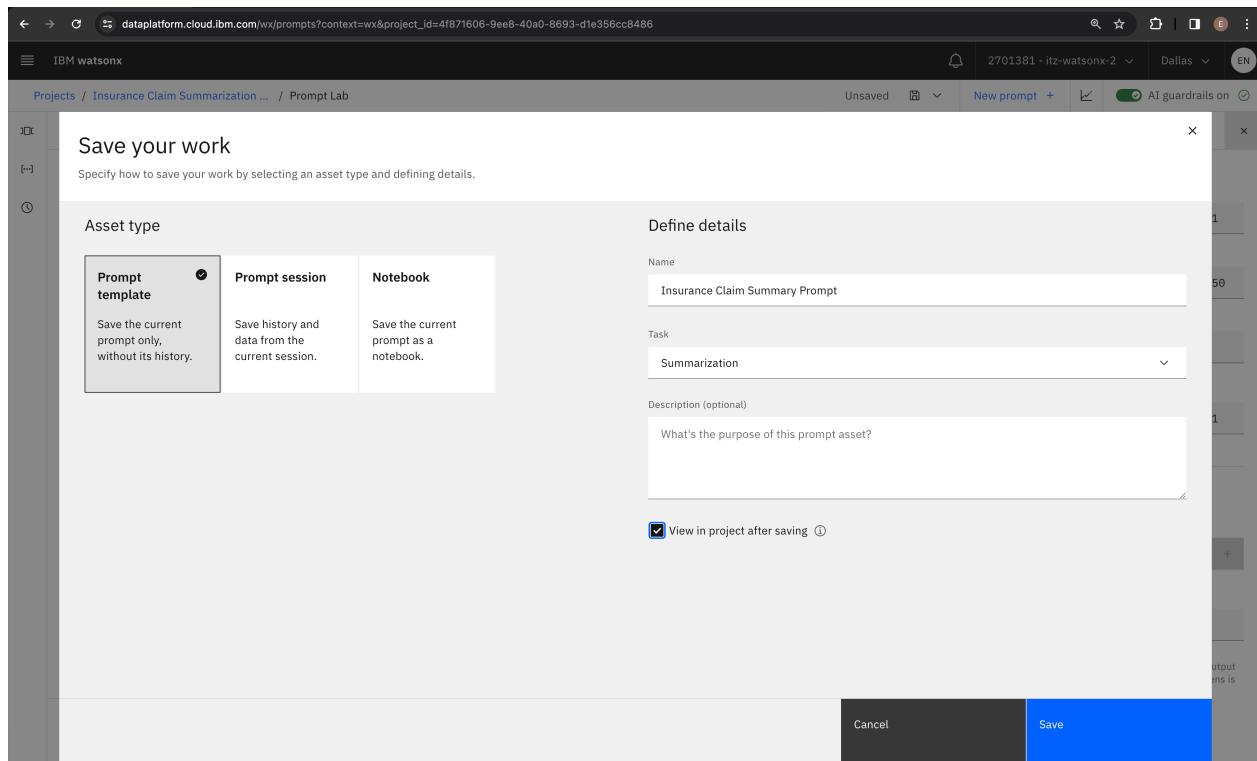
- Repetition penalty:** Set to 1.
- Stopping criteria:** Set to 'Stop sequences'.
- Min tokens:** Set to 0.
- Max tokens:** Set to 200 (highlighted with a blue border).
- Note:** A note below says 'Enter up to 6 sequences to stop output after the minimum number of tokens is reached.'
- Buttons:** 'Generate' and 'Reset to default'.

1.3 Saving Prompt

Setelah semuanya selesai, klik ikon **Save** di bagian kanan atas, lalu klik **Save as**. Ini akan membawa Anda ke layar berikutnya, tempat Anda akan memilih jenis aset, memberi nama *prompt template*, dan memilih klasifikasi dari *prompt* yang dibuat. Pastikan untuk memilih opsi "Prompt template", lalu isi nama dengan cara yang sama seperti yang ditunjukkan pada gambar. Dari sana, pilih **Summarization** sebagai nilai *Task*, dan pastikan untuk memilih **View in project after saving**.

The screenshot shows the 'Save as' dialog in the WatsonX Prompt Lab. It includes fields for 'Save as' name, 'Model' (set to 'llama-2-70b-chat'), and 'Task' (set to 'Summarization'). On the right, there are 'Model parameters' and 'Stopping criteria' sections:

- Model parameters:**
 - top p (nucleus sampling):** Set to 1.
 - Top K:** Set to 50.
 - Random seed:** An empty input field.
 - Repetition penalty:** Set to 1.
- Stopping criteria:** Set to 'Stop sequences'.



Super! Sekarang Anda seharusnya melihat aset tersebut pada layar saat Anda kembali ke halaman Aset. Pastikan Anda telah menyelesaikan langkah sebelumnya dengan benar dengan mengeklik 3 titik di sebelah kanan aset yang baru terbuat dan pilih **Evaluate**. Kita akan menguji *prompt* baru kita!

2. Mengevaluasi Prompt

Anda akan melihat layar baru dengan tombol **Evaluate** di bagian tengah layar. Klik tombol tersebut.

Run a new evaluation

Run an evaluation job

Click Evaluate to choose dimensions to evaluate and select test data.

Evaluate

Sekarang Anda akan melihat layar yang menanyakan dimensi apa yang akan dievaluasi. Pada tahap ini, kita hanya dapat mengevaluasi kualitas GenAI (kesehatan model diaktifkan secara default).

Evaluate prompt template

Choose the evaluation dimensions and select the test data. [Learn more](#)

<input checked="" type="checkbox"/> Dimension	Description	Show less ^
<input checked="" type="checkbox"/> Generative AI Quality	The Generative AI Quality monitor calculates a variety of metrics based on prompt template task type. Some metrics compare model output to the reference output you provide. Other metrics analyze model input and output and do not require reference output.	Advanced settings

[Cancel](#) [Back](#) [Next](#)

Sekarang, kita akan gunakan kedua data berikut: [summarization validation data](#) dan [summarization test data](#) yang ada di dalam folder "data" pada github ini. Untuk sekarang, kita akan gunakan *validation data*. Anda dapat menggunakan *test data* untuk melakukan [Step 6](#).

Evaluate prompt template

Choose the evaluation dimensions and select the test data. [Learn more](#)

Select dimensions

Select test data

Review and evaluate

Drop a file here or browse for a file to upload

Add a CSV file that include input and output examples. Maximum size is 8 MB. Maximum number of records is 1000. Minimum number of records is 10.

Browse

Cancel Back Next

Sekarang saatnya memetakan variabel masukan ke variabel *prompt* yang telah kita buat sebelumnya. Karena kita menggunakan file CSV, kita pilih koma sebagai *delimiter*. Selain itu, kita pilih "input" untuk dipetakan ke kolom "Insurance_Claim" di dalam file CSV. Terakhir, "Reference output" adalah kolom pada dataset yang menyimpan keluaran dari LLM. Ini adalah cara kita mengukur seberapa efektif kita telah menyusun prompt kita.

Evaluate prompt template

Choose the evaluation dimensions and select the test data. [Learn more](#)

Select dimensions

Select test data

Review and evaluate

For each prompt variable, select the associated column. [Learn more](#)

Field separation

Select delimiter

Comma (,)

Input

input

Insurance_Claim

Reference output

Reference output

Summary

Cancel Back Next

Setelah itu, Anda akan melihat layar untuk meninjau informasi yang telah Anda masukkan. Setelah Anda memverifikasi bahwa Anda telah memasukkan semuanya dengan benar, tekan tombol "Evaluate". dan tunggu hingga proses evaluasi selesai. Proses ini dapat memakan waktu hingga beberapa menit.

Evaluate prompt template

Choose the evaluation dimensions and select the test data. [Learn more](#)

Select dimensions	<input checked="" type="radio"/>	Review
Select test data	<input checked="" type="radio"/>	Task: Text summarization
Review and evaluate	<input checked="" type="radio"/>	Test data: Insurance claim summarization validation data.csv
Evaluations: Generative AI Quality		

Note:
Evaluation can take a few minutes to complete. You can continue to work on other things while your evaluation is in progress.

[Cancel](#) [Back](#) [Evaluate](#)

Setelah evaluasi selesai, Anda akan melihat layar yang menampilkan hasilnya. Seperti yang dapat Anda lihat di sini, evaluasi kami tidak berhasil, memicu 13 peringatan dan gagal dalam pengujiannya. Ini sudah diduga! Sasaran lab ini adalah untuk membuat *prompt* pertama, dan berupaya mencapai hasil yang lebih baik sehingga kita dapat menemukan *prompt* yang baik dan akhirnya membawanya ke tahap produksi.

Last evaluation: Wed, Dec 6, 2023, 1:44 PM PST

[Actions](#)

Deployment details	Test details	Model health →
Test data set Insurance claim summarization validation data.csv	1 Tests run Tests passed: 0 Tests failed: 1	Records ① 10 Records Latency (record) ① 2,674 ms Median record latency Token count ① 1,938 Total input token count 326 Total output token count

Generative AI Quality - Text summarization

Alerts triggered: 13

[Feedback](#)

3. Cara Alternatif untuk Melakukan Evaluasi Prompt

Bagi yang tertarik, ada cara lain untuk menjalankan evaluasi dari dalam sesi prompt lab. Klik ikon di kanan atas yang seperti grafik, yang merupakan tombol untuk mengevaluasi.

The screenshot shows the IBM Watsonx Prompt Lab interface. The 'Structured' tab is selected. The 'Set up' section contains an optional instruction and examples. The 'Try' section shows a test input '{input}' and its generated output 'Generated output appears here.'. The 'Evaluate' section on the right contains various model parameters and stopping criteria, with a prominent blue 'Generate' button.

Jika Anda belum melakukannya, Anda akan diminta untuk menyimpan pekerjaan Anda dengan cara yang sama seperti sebelumnya. Kemudian, Anda akan langsung diarahkan ke halaman yang meminta Anda untuk memilih dimensi yang akan dievaluasi, dan melanjutkan dengan proses yang sama seperti [Step 2](#).

The screenshot shows the 'Evaluate prompt template' step in the Watsonx interface. The 'Select dimensions' tab is selected. A table lists evaluation dimensions, with 'Generative AI Quality' checked. The 'Next' button is highlighted in blue at the bottom.

4. Hasil Evaluasi

Last evaluation: Wed, Dec 6, 2023, 1:44 PM PST

Deployment details

Test data set
Insurance claim summarization validation data.csv

Test details

1 Tests run

Tests passed: 0, Tests failed: 1

Model health

Records: 10 Records

Latency (record): 2,674 ms Median record latency

Token count: 1,938 Total input token count 326 Total output token count

Generative AI Quality - Text summarization

Alerts triggered: 13

Feedback

Gulir ke bawah halaman **Evaluate** untuk melihat metrik spesifik yang digunakan untuk mengevaluasi prompt Anda. LLM dapat dievaluasi menggunakan berbagai metrik, dan metrik yang digunakan bergantung pada tugas yang diminta untuk diselesaikan oleh LLM Anda.

Metric	Score	Violation
ROUGE-1	0.32	0.48
ROUGE-2	0.16	0.64
ROUGE-L	0.29	0.51
ROUGE-Lsum	0.30	0.50
SARI	36.49	43.51
METEOR	0.34	0.46
F1 Score	0.29	0.51
Precision	0.26	0.54
Recall	0.40	0.40
BLEU	0.13	0.67
Jaccard similarity	0.17	0.63

Dalam lab ini, tugas LLM adalah melakukan **Summarization** yang dapat diukur menggunakan berbagai metrik kualitas. Kami menyarankan Anda meluangkan waktu sekarang untuk meninjau dokumentasi tentang [metrik kualitas AI generatif yang didukung oleh watsonx.governance](#).

Supported generative AI quality metrics

The following generative AI quality metrics are supported by watsonx.governance:

- ✓ ROUGE
- ✓ SARI
- ✓ METEOR
- ✓ Text quality
- ✓ BLEU
- ✓ Sentence similarity
- ✓ PII
- ✓ HAP
- ✓ Readability
- ✓ Exact match
- ✓ Multi-label/class metrics

Setelah membaca deskripsi metrik evaluasi LLM tersebut, Anda mungkin menyadari bahwa tampilan metrik bawaan pada halaman **Evaluation** tidak terlalu membantu. Jangan khawatir, Anda akan belajar cara mengakses representasi metrik yang lebih intuitif di **AI Factsheet**.

Di sudut kiri atas layar, pilih tab **AI Factsheet**, yang akan menampilkan rincian hasil yang lebih jelas. Hal pertama yang akan Anda lihat di halaman ini adalah pilihan untuk melacak prompt Anda dalam sebuah AI use case. Kami akan membahas opsi ini di lab berikutnya.

The screenshot shows the IBM WatsonX interface with the URL dataplatform.cloud.ibm.com/wx/prompt-details/0f37828d-0ba6-4ba3-afd4-ee3d349d8b49/factsheet?context=wx&project_id=4d0f001c-b5.... The top navigation bar includes links for Upgrade, Andrew Bruneel's Account, Dallas, and AB. The left sidebar has tabs for AI Factsheet (selected), Evaluate, Governance, Foundation model, Prompt template, Prompt parameters, Evaluation, Develop, Test, Attachments, and Other attachments. The main content area is titled 'Governance' and shows a message: 'This prompt template is not tracked. To track a prompt template, add it to an AI use case. Tracking capture details about the asset through its lifecycle as part of governance strategy.' A blue button labeled 'Track in AI use case' is present. The bottom of the sidebar also lists 'Foundation model'.

Gulir ke bawah halaman ke bagian **Generative AI Quality**, dan Anda akan melihat visual yang mendampingi metrik yang kita lihat sebelumnya. Di sini, Anda bisa melihat seberapa dekat prompt kita dengan standar kelulusan uji. Secara keseluruhan, tampaknya masih banyak yang harus kita kerjakan!

The screenshot shows the AI Factsheet section of the IBM WatsonX platform. On the left, a sidebar lists navigation options: Governance, Foundation model, Prompt template, Prompt parameters, Evaluation, Develop, and Test (which is selected). The main area displays 'Threshold alerts' with a count of 13. Below this is a section titled 'Generative AI Quality' containing six horizontal sliders. The first slider for 'Readability' has a value of 56.97. The second slider for 'Recall' has a value of 0.40. The third slider for 'Input data HAP' has a value of 0. The fourth slider for 'ROUGE-2' has a value of 0.16. The fifth slider for 'Jaccard similarity' has a value of 0.17. The sixth slider for 'METEOR' has a value of 0.34. The seventh slider for 'ROUGE-LSum' has a value of 0.30. The eighth slider for 'Output data HAP' has a value of 0.

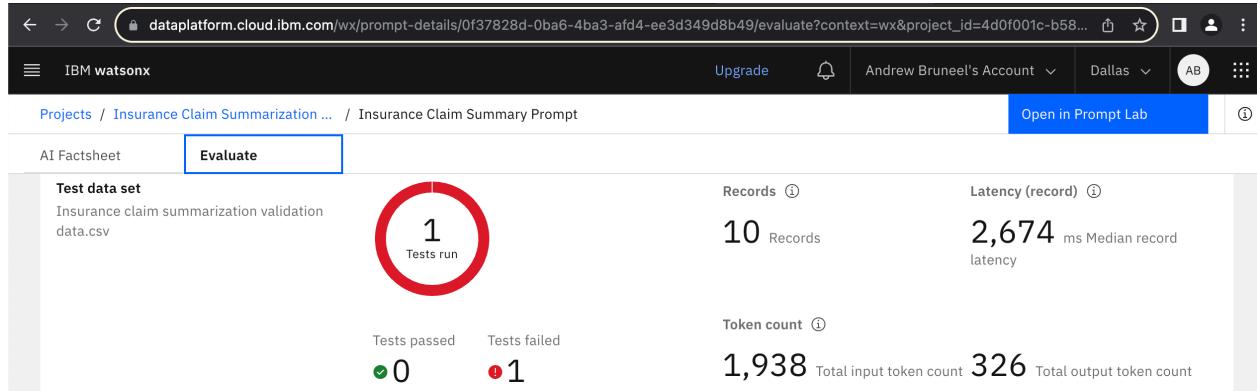
Dengan **AI Factsheet** terbuka, luangkan waktu untuk membaca [metrik kualitas AI generatif yang didukung oleh watsonx.governance](#). Dengan membandingkan **AI Fact Sheet**, Anda dapat melihat bahwa setiap metrik memiliki batas atas dan batas bawah, dengan rentang ideal ditunjukkan oleh bagian yang lebih tebal di sisi kiri dari setiap pengukuran metrik.

Segitiga terbalik menunjukkan nilai yang diukur untuk setiap metrik. Segitiga hitam menunjukkan hasil baik, sedangkan segitiga merah menunjukkan hasil buruk. Perhatikan baik-baik, dan Anda akan melihat bahwa evaluasi Anda hanya memiliki tiga metrik yang memenuhi standar.



Setiap metrik yang mengukur kualitas ringkasan berada di zona merah. Semoga Anda bisa menghadapi tantangan ini dengan memilih salah satu metrik untuk ditingkatkan dan membangun prompt yang lebih baik untuk kasus meringkas klaim asuransi ini.

Namun, untuk saat ini, saya akan menunjukkan cara memodifikasi ambang batas untuk metrik ini sehingga prompt Anda memiliki persyaratan yang lebih longgar untuk lulus pengujian. Dalam penggunaan dunia nyata, hal ini hanya boleh dilakukan jika benar-benar diperlukan dan sesuai dengan kebutuhan Anda. Kembali ke layar evaluasi, Anda akan melihat tombol biru di sebelah kanan "Generative AI Quality - Text Summarization" di mana kita dapat menyesuaikan ambang batas kita.



The screenshot shows the AI Factsheet for the Insurance Claim Summary Prompt. It includes sections for Test data set, Records, Latency, Token count, and a detailed view of Generative AI Quality metrics (ROUGE-1, ROUGE-2) with an alert count of 13. A red circle highlights the 'Evaluate' button in the top navigation bar.

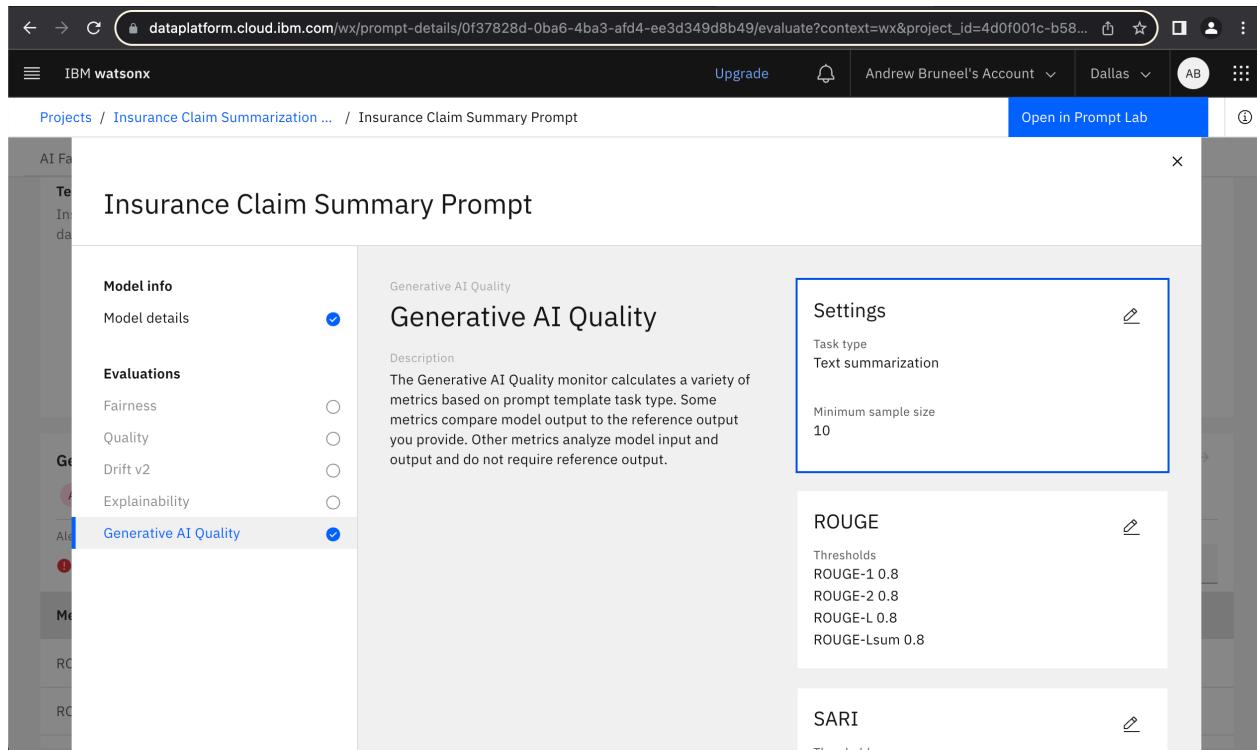
Metric	Score	Violation
ROUGE-1	0.32	0.48
ROUGE-2	0.16	0.64

Generative AI Quality - Text summarization

Alerts triggered: 13

Feedback

Silakan klik tombol tersebut, dan Anda akan melihat layar berikut:



The screenshot shows the configuration panel for Generative AI Quality. The 'Generative AI Quality' tab is selected in the sidebar. The main area displays the description of the monitor and its settings. The 'Settings' section shows a minimum sample size of 10. Below it are sections for ROUGE and SARI, each with their respective threshold configurations.

Generative AI Quality

Description

The Generative AI Quality monitor calculates a variety of metrics based on prompt template task type. Some metrics compare model output to the reference output you provide. Other metrics analyze model input and output and do not require reference output.

Settings

Task type: Text summarization
Minimum sample size: 10

ROUGE

Thresholds:
ROUGE-1: 0.8
ROUGE-2: 0.8
ROUGE-L: 0.8
ROUGE-Lsum: 0.8

SARI

Di sini, kita akan menyesuaikan metrik keterbacaan (readability). Metrik ini akan lebih tinggi jika respons yang dihasilkan model terhadap prompt Anda mudah dipahami, dan lebih rendah jika sulit untuk dibaca. Gulir ke bawah halaman dan klik tombol edit untuk memodifikasinya.

The screenshot shows the IBM Watsonx interface with the URL https://dataplatform.cloud.ibm.com/wx/prompt-details/0f37828d-0ba6-4ba3-afd4-ee3d349d8b49/evaluate?context=wx&project_id=4d0f001c-b58.... The page title is "Insurance Claim Summary Prompt". On the left sidebar, under "Evaluations", "Generative AI Quality" is selected. In the main panel, there are two sections: "Output data HAP" and "Input data HAP", both showing thresholds at 0%. Below them is a section for "Readability" with a threshold of 60. An "Edit" button is visible next to the readability section.

Terakhir, silakan ubah keterbacaan sedikit, misalnya menjadi 55, yang menempatkan persyaratan tersebut dalam rentang "Cukup sulit untuk dibaca." Setelah itu, Anda dapat mengklik **save** untuk menyelesaikan perubahan Anda. Sebagai pengingat, metrik ini tidak boleh diubah terlalu banyak dalam penggunaan dunia nyata!

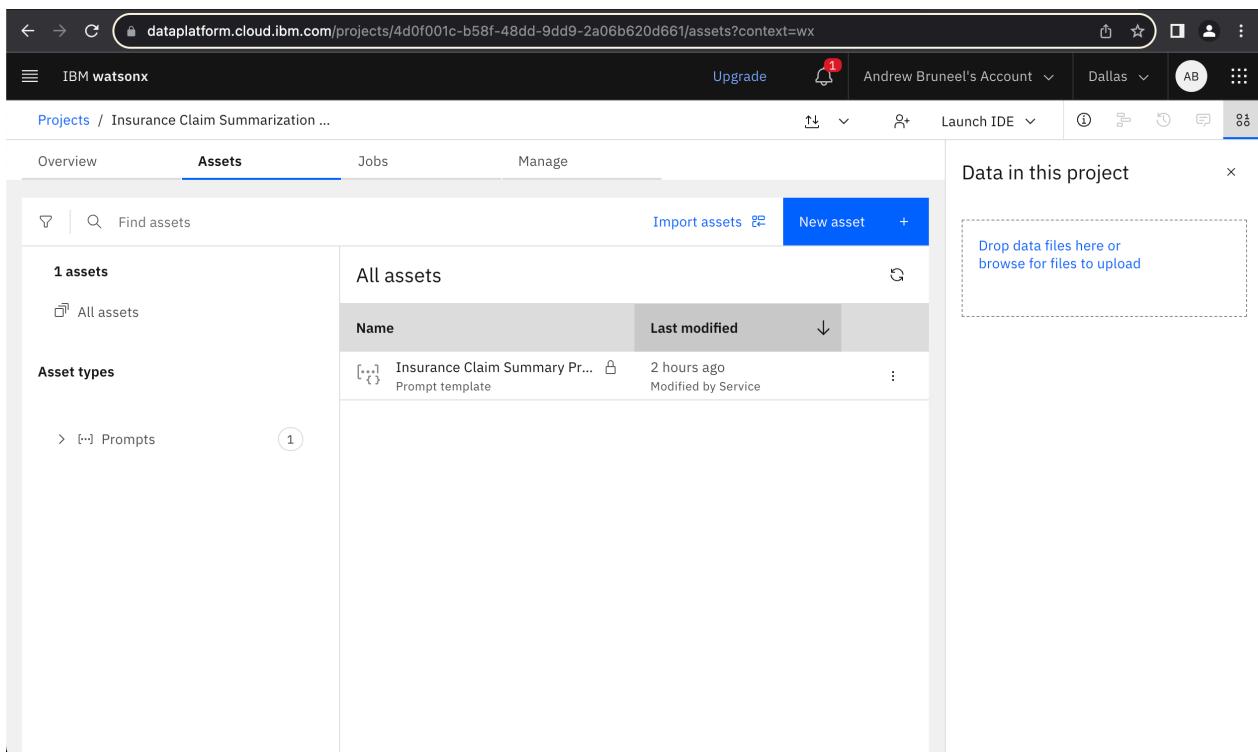
The screenshot shows the "Readability" configuration dialog from the IBM Watsonx interface. The "Lower thresholds" field contains the value 55. At the bottom right, there are "Cancel" and "Save" buttons, with the "Save" button highlighted by a blue box.

****Catatan: Seilahkan merujuk ke [powerpoint ini](#) untuk informasi lebih lanjut tentang setiap metrics!**

5. Percobaan dengan Prompt Baru

Sekarang Anda telah melihat cara membangun template prompt Anda sendiri dan mengevaluasinya terhadap dataset pengujian, saatnya untuk bereksperimen! Tujuan Anda adalah membuat template prompt yang akan lulus tes terhadap dataset validasi yang diberikan kepada Anda di lab.

Pertama, mari kita mulai dengan membuat dua salinan dari template prompt yang ada yang dapat Anda gunakan untuk dimodifikasi dan dieksplorasi. Klik "Insurance Claim Summarization..." di sudut kiri atas layar Anda, yaitu hyperlink setelah "Projects". Ini akan membawa Anda kembali ke halaman "Assets", yang akan terlihat seperti ini:



The screenshot shows the IBM WatsonX interface on a web browser. The URL in the address bar is dataplatform.cloud.ibm.com/projects/4d0f001c-b58f-48dd-9dd9-2a06b620d661/assets?context=wx. The top navigation bar includes links for Upgrade, Andrew Bruneel's Account, Dallas, and AB. Below the header, there are tabs for Projects, Assets (which is selected), Jobs, and Manage. The Assets tab has sub-options for Overview, Import assets, New asset, and a search bar labeled 'Find assets'. On the left, there's a sidebar with sections for Asset types (Prompts) and a list of 1 asset. The main content area displays a table titled 'All assets' with columns for Name and Last modified. One asset is listed: 'Insurance Claim Summary Pr...' (Prompt template), last modified 2 hours ago by Service. To the right of the table is a sidebar titled 'Data in this project' with a section for 'Drop data files here or browse for files to upload'.

5.1 Memakai Granite

Klik kembali ke prompt Anda yang ada, kita akan mengubah LLM yang digunakan untuk menghasilkan ringkasan untuk kita. Setelah Anda melihat layar prompt kembali, matikan pengaturan "Autosave" di bagian header halaman. Kemudian, klik dropdown pemilihan model dan pilih "View all foundational models".

The screenshot shows the IBM Watsonx Prompt Lab interface. On the left, there's a sidebar with sections like 'Set up' and 'Examples (optional)'. In the center, there's a 'Try' section where you can test your prompt. On the right, there's a sidebar for 'Model: llama-2-70b-chat' which includes a dropdown for 'Recent models' (llama-2-70b-chat, flan-ul2-20b, granite-13b-chat-v2, starcoder-15.5b), 'Stopping criteria' (with 'Stop sequences' and token limits), and a 'Reset to default' button.

Pilih IBM's granite-13b-chat-v2 sebagai model untuk prompt pembuat ringkasan ini.

The screenshot shows the 'Select a foundation model' dialog box. It lists several models in a grid format:

Model	Provider	Source	Provider	Source	Provider	Source	Provider	Source
flan-t5-xl-3b	Google	Hugging Face	flan-t5-xxl-11b	BigCode	Hugging Face	granite-13b-chat-v1	BigScience	Hugging Face
granite-13b-instruct-v1	Google	Hugging Face	granite-13b-instruct-v2	IBM	Hugging Face	mpt-7b-instruct2	EleutherAI	Hugging Face

Each row contains a thumbnail icon, the model name, provider, and source information. The 'granite-13b-chat-v2' model is highlighted in blue, indicating it has been selected.

Anda akan melihat halaman muncul yang menampilkan informasi berguna tentang model yang tersedia, serta tombol biru "Select Model". Silakan klik tombol tersebut setelah Anda siap.

Model Datasheet

Granite Base 13 Billion Model Chat (granite.13b.chat.v2) Details

IBM Generative AI Large Language Foundation Models are Enterprise-level English-language models trained with large a volume of data that has been subjected to intensive pre-processing and careful analysis. The Granite 13 Billion chat V2.0 (granite.13b.chat.v2) model is the chat-focused variant initialized from the pre-trained Granite Base 13 Billion Base V2.0 (granite.13b.base.v2) model. granite.13b.base.v2 has been trained using over 2.5T tokens. The Granite family of models will support all 5 language tasks (Q&A, Generate, Extract, Summarize, and Classify).

granite.13b.chat.v2 is an chat-focused model that was tuned to improve its ability to perform Retrieval Augmented Generation (RAG) use cases. The team applied a two step process of Supervised Fine Tuning (SFT) and IBM's novel Reinforcement Learning with AI Feedback (RLAIF) technique, [Salmon](#).

Back Select model →

Dari sini, kita akan menggunakan opsi "Save as" sekali lagi, yang akan membuat sebuah instance baru dari template prompt di halaman aset kita.

Save as

Model parameters

Decoding
Greedy Sampling

Repetition penalty
1 2 1

Stopping criteria

Stop sequences
Min tokens 0 Max tokens 200

Enter up to 6 sequences to stop output after the minimum number of tokens is reached.

Reset to default

Generate →

Isi informasi di halaman "Save Your Work" seperti yang saya contohkan di bawah ini. Pastikan untuk memasukkan nama model Anda di dalam kolom "Name", karena ini akan penting saat kita membuat template prompt berikutnya!

Save your work

Specify how to save your work by selecting an asset type and defining details.

Asset type

- Prompt template** (selected)
- Prompt session**
- Notebook**

Define details

Name: Summary - granite-13b-chat-v2

Task: Summarization

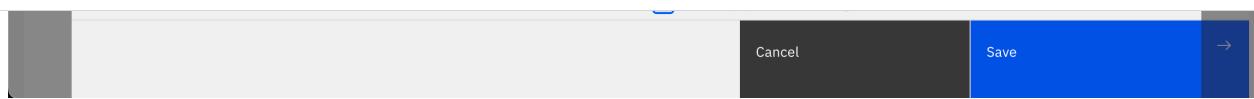
Description (optional): What's the purpose of this prompt asset?

Only the last prompt test is included in the prompt template

workshop-telkom-partnership-feb2025

/ Lab 6 (Optional) - Evaluate prompt using watsonx.governance /

Top ↑



kerja bagus -- sekarang kita seharusnya bisa melihat template baru di dalam halaman Project, sekaligus dengan template asli yang kita buat sebelumnya.

Projects / Insurance Claim Summarization ...

Overview Assets Jobs Manage

Find assets Import assets New asset

2 assets All assets

All assets

Name	Last modified
Summary - granite-13b-chat-v2	Now Modified by you
Insurance Claim Summary Pr...	Now Modified by Service

Asset types

Prompts

Data in this project

Drop data files here or browse for files to upload

5.2 Cara Lainnya!

Sekarang setelah Anda melihat betapa cepatnya membuat salinan dari template prompt asli Anda, Anda dapat membuat template baru ke model yang Anda pilih! Ulangi langkah yang Anda lakukan untuk membuat template prompt yang dievaluasi dengan granite, tetapi kali ini pilih model baru (Pastikan untuk mengklik ke **template asli** yang kami buat, bukan template granite). Model tersebut bisa apa saja yang Anda inginkan! Pastikan untuk **menyertakan nama model** dalam penamaan template prompt baru Anda, seperti yang kita lakukan dengan granite. Ini akan membantu Anda dengan cepat membedakan antara template prompt di kemudian hari. Berikut adalah tab Aset saya setelah saya menambahkan satu model lagi:

Name	Last modified	Actions
[...] Summary - llama-2-70b-chat Prompt template	Now Modified by you	[...]
[...] Insurance Claim Summary Pr... Prompt template	Now Modified by Service	[...]
[...] Summary - granite-13b-chat-v2 Prompt template	20 minutes ago Modified by Service	[...]

6. Percobaan dengan Seting

Sekarang setelah Anda mencapai tahap ini, Anda siap untuk bereksperimen dan meningkatkan kinerja prompt Anda! Tujuannya adalah agar Anda dapat lulus evaluasi terhadap data validasi menggunakan alat yang tersedia untuk Anda. Silakan melakukan perubahan sebanyak yang Anda mau, termasuk:

- Membuat template prompt baru dengan model yang berbeda
- Mengubah instruksi prompt
- Menyediakan contoh (few-shot prompting)
- Mengubah parameter model seperti:
 - Greedy vs. Sampling
 - Temperature
 - Min/Max Tokens
 - Top P/Top K
 - Stop Sequence

- Mengevaluasi template prompt menggunakan dataset yang berbeda

Dan perubahan lain yang Anda pikirkan! Mungkin juga berguna sebagai langkah awal bagi Anda untuk melihat lebih dekat pada data validasi. Dari sana, Anda dapat memutuskan modifikasi mana yang akan difokuskan yang menurut Anda akan paling efektif dalam meningkatkan hasil Anda. Pastikan untuk merujuk pada langkah-langkah sebelumnya dari lab jika Anda memiliki pertanyaan; langkah-langkah tersebut seharusnya dapat menjawab sebagian besar pertanyaan yang Anda miliki!

****Catatan: Ingat bahwa jika Anda mengalami kesulitan, Anda dapat memodifikasi ambang batas untuk evaluasi Anda jika diperlukan.** Beberapa ambang batas memiliki nilai default yang tinggi dan sulit dicapai bahkan oleh model-model mutakhir, jadi jika Anda mendapatkan diri Anda terus-menerus berada jauh di bawah nilai ambang batas meskipun telah mencoba solusi di atas, pertimbangkan untuk memodifikasi nilai ambang batas Anda. PowerPoint yang dibagikan dalam sumber daya lab ini adalah sumber yang baik untuk melihat metrik mana yang mungkin berguna untuk menurunkan ambang batas.

Contoh few-shot prompting:

Di bawah ini adalah contoh singkat tentang penggunaan few-shot prompting dalam konteks data yang digunakan untuk lab ini. Pastikan bahwa saat Anda memutuskan untuk menggunakan few-shot prompting, Anda menggunakan contoh yang **berbeda** dari apa yang ada di data validasi dan pengujian Anda. Jika tidak, Anda akan melatih model Anda pada skenario yang tepat yang akan dilihat dalam data pengujian, yang bukanlah skenario realistik yang akan Anda temui dalam penggunaan dunia nyata! Contoh ini menunjukkan satu shot, tetapi silakan tambahkan lebih banyak untuk membuat model Anda lebih terfokus pada contoh yang Anda berikan!

dataplatform.cloud.ibm.com/wx/prompts/templates/898fbeca-1702-4415-9cbf-f8e30957e7a2?project_id=4d0f001c-b58f-48dd-9dd9-2a06...

IBM watsonx

Upgrade Andrew Bruneel's Account Dallas AB :

Projects / Insurance Claim Summarization ... /

AI guardrails on Autosave on

Summary - llama-2-70b-chat

Prompt: Autosaved 10:06 AM

Evaluate New prompt + Save work

Structured Freeform

Model: llama-2-70b-chat

Examples (optional)

Input:

On December 8th, 2022, at 1:00 AM, my vehicle, a Toyota Corolla, was involved in a severe accident in New Hampshire. Our driver was traveling 10mph over the speed limit when he collided with another car, causing a frontal collision. The impact was significant, resulting in extensive damage to both vehicles. My driver and I immediately contacted emergency services and received medical attention at the scene, as well as the other driver. My vehicle sustained damage to the front bumper, hood, and windshield. My driver and I suffered mild injuries and are receiving ongoing medical treatment. I immediately called your agent and submitted the claim form giving the accident details. I am also supporting witness statements and photographs of the damaged vehicle.

Output:

My vehicle was involved in a severe accident in New Hampshire. The vehicle sustained damage to the hood, front bumper, and windshield.

Add example +

Generate →

Saat Anda terus melakukan perubahan, prompt juara Anda dengan kinerja terbaik akan dipilih untuk digunakan di lab berikutnya!

