

Starting April 29, 2025, Gemini 1.5 Pro and Gemini 1.5 Flash models are not available in projects that have no prior usage of these models, including new projects. For details, see [Model versions and lifecycle](#) (/vertex-ai/generative-ai/docs/learn/model-versions#legacy-stable).

# Create MedLM prompts

**Caution:** MedLM is deprecated. Access to MedLM will no longer be available on or after September 29, 2025.

The available MedLM models, MedLM-medium and MedLM-large, are foundation models for medical question answering and summarization. You can access the models using the Vertex AI MedLM API. This page gives you an overview of the available MedLM models, the APIs you use to interact with the models, and ways to customize their behaviors.

## Before you begin

- See [MedLM models overview](#) (/vertex-ai/generative-ai/docs/medlm/overview) for information including customer responsibilities, regulatory information, and [Responsible AI](#) (/vertex-ai/generative-ai/docs/learn/responsible-ai) best practices.
- See the MedLM model card for model details, such as MedLM's intended use, data overview, and safety information. Click the following link to download a PDF version of the MedLM model card:

↓ [Download the MedLM model card](#)

(/static/vertex-ai/generative-ai/docs/medlm/MedLM-model-card.pdf)

## Prompt design

To interact with the MedLM models, you send natural language instructions, also called prompts, that tell the model what you want it to generate. However, LLMs can sometimes behave in unpredictable ways. Prompt design is an iterative process of trial and error that takes time and practice to become proficient in. To learn about general prompt design strategies, see [Introduction to prompt design](#) (/vertex-ai/generative-ai/docs/learn/introduction-prompt-design). For task-specific prompt

design guidance for text, see [Overview of prompting strategies](#) ([/vertex-ai/generative-ai/docs/learn/prompts/prompt-design-strategies](#)).

## Use cases

- **Summarization:** Create a shorter version of a document that incorporates pertinent information from the original text. For example, you might want to summarize a medical note describing an outpatient visit, and to extract relevant information for specific data points.
- **Question answering:** Provide answers to questions in text. For example, you might want to ask a general medical question to generate answers from the knowledge base.

## Supported models

- `medlm-medium`
- `medlm-large`

## Get started

The following samples show how to get started with the MedLM API using the following interfaces:

- The Vertex AI REST API
- Vertex AI SDK for Python
- Vertex AI Studio

[RESTPython \(Colaboratory\)...](#)  
(#rest)


[Vertex AI Studio...](#)



**Caution:** The generated output may not always be completely reliable. Due to the nature of LLMs and Generative AI, outputs may have incorrect or biased information and should be reviewed. All summaries or answers should be considered draft and not final.

Before using any of the request data, make the following replacements:

- **`PROJECT_ID`** : Your [project ID](#) ([/resource-manager/docs/creating-managing-projects#identifiers](#)).

- MEDLM\_MODEL : The MedLM model, either medlm-medium or medlm-large.

HTTP method and URL:

POST [https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT\\_ID/locations/region/publishers/google/models/MEDLM\\_MODEL:predict](https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT_ID/locations/region/publishers/google/models/MEDLM_MODEL:predict)

Request JSON body:

```
{
  "instances": [
    {
      "content": "Question: What causes you to get ringworm?"
    }
  ],
  "parameters": {
    "temperature": 0,
    "maxOutputTokens": 256,
    "topK": 40,
    "topP": 0.95
  }
}
```

To send your request, choose one of these options:

curlPowerShell (#powershell)  
(#curl)

★ **Note:** The following command assumes that you have logged in to the **gcloud** CLI with your user account by running **gcloud init** (/sdk/gcloud/reference/init) or **gcloud auth login** (/sdk/gcloud/reference/auth/login) , or by using **Cloud Shell** (/shell/docs), which automatically logs you into the **gcloud** CLI . You can check the currently active account by running **gcloud auth list** (/sdk/gcloud/reference/auth/list).

Save the request body in a file named `request.json`. Run the following command in the terminal to create or overwrite this file in the current directory:

```
cat > request.json << 'EOF'
{
  "instances": [
    {
      "content": "Question: What causes you to get ringworm?"
    }
  ],
  "parameters": {
    "temperature": 0,
    "maxOutputTokens": 256,
    "topK": 40,
    "topP": 0.95
  }
}
EOF
```

Then execute the following command to send your REST request:

```
curl -X POST \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  -H "Content-Type: application/json; charset=utf-8" \
  -d @request.json \
  "https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT_ID
```

## Question answering prompts



The following sections contain question answering prompt samples. Each sample prompt includes the recommended model and parameter values.

### Long-form question answering

The following samples show how the MedLM API answers a long-form medical question formulated as a query.

**Caution:** The generated output may not always be completely reliable. Due to the nature of LLMs and Generative AI, outputs may have incorrect or biased information and should be reviewed. All summaries or answers should be considered draft and not final.

Before using any of the request data, make the following replacements:

- PROJECT\_ID : Your project ID (/resource-manager/docs/creating-managing-projects#identifiers).
- MEDLM\_MODEL : The MedLM model, either `medlm-medium` or `medlm-large`.

HTTP method and URL:

POST `https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT_ID/loc`

Request JSON body:

```
{
  "instances": [
    {
      "content": "Question: What causes you to get ringworm?"
    }
  ],
  "parameters": {
    "temperature": 0,
    "maxOutputTokens": 256,
    "topK": 40,
    "topP": 0.95
  }
}
```

To send your request, choose one of these options:

```
curlPowerShell (#powershell)
(#curl)
```

★ **Note:** The following command assumes that you have logged in to the **gcloud** CLI with your user account by running **`gcloud init`** (/sdk/gcloud/reference/init) or **`gcloud auth login`**

(/sdk/gcloud/reference/auth/login) , or by using [Cloud Shell](#) (/shell/docs), which automatically logs you into the **gcloud** CLI . You can check the currently active account by running **[gcloud auth list](#)** (/sdk/gcloud/reference/auth/list).

Save the request body in a file named `request.json`. Run the following command in the terminal to create or overwrite this file in the current directory:

```
cat > request.json << 'EOF'
{
  "instances": [
    {
      "content": "Question: What causes you to get ringworm?"
    }
  ],
  "parameters": {
    "temperature": 0,
    "maxOutputTokens": 256,
    "topK": 40,
    "topP": 0.95
  }
}
EOF
```

Then execute the following command to send your REST request:

```
curl -X POST \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  -H "Content-Type: application/json; charset=utf-8" \
  -d @request.json \
  "https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT_ID
```

**Caution:** The generated output may not always be completely reliable. Due to the nature of LLMs and Generative AI, outputs may have incorrect or biased information and should be reviewed. All summaries or answers should be considered draft and not final.

## Multiple choice question answering

The following samples show how the MedLM API answers a multiple choice medical question. The prompt is the following:

Instructions: This text contains multiple-choice questions about medical knowledge. S

Question 1: Which medication causes the maximum increase in prolactin level?

- (A) Risperidone
- (B) Clozapine
- (C) Olanzapine
- (D) Aripiprazole

Explanation: To solve this question, let's refer to authoritative sources. Clozapine

Answer: (A)

Question 2: What is the recommended age for routine screening mammography?

- (A) 20 years
- (B) 30 years
- (C) 40 years
- (D) 50 years

Explanation: The age of routine screening may vary depending on the country. In the U

Answer: (C)

Question 3: A 65-year-old male experiences severe back pain and paralysis in his left

- (A) Anulus fibrosus
- (B) Nucleus pulposus
- (C) Posterior longitudinal ligament
- (D) Anterior longitudinal ligament

Explanation: This man's symptoms and imaging findings are consistent with a herniated

Answer: (B)

Question 4: Which cells in the lungs are also known as APUD cells?

- (A) Dendritic cells
- (B) Type I pneumocytes
- (C) Type II pneumocytes
- (D) Neuroendocrine cells

Explanation: Neuroendocrine cells, also known as Kultschitsky-type cells, Feyrter cel

Answer: (D)

Question 5: Which microorganism indicates remote contamination of water?

- (A) Streptococci
- (B) Staphylococci
- (C) Clostridium perfringens
- (D) Vibrio

Explanation: The presence of Clostridium perfringens in water indicates remote contamination.



Answer: (C)

## REST

(#rest)

**Caution:** The generated output may not always be completely reliable. Due to the nature of LLMs and Generative AI, outputs may have incorrect or biased information and should be reviewed. All summaries or answers should be considered draft and not final.

Before using any of the request data, make the following replacements:

- PROJECT\_ID : Your project ID (/resource-manager/docs/creating-managing-projects#identifiers).
- MEDLM\_MODEL : The MedLM model, either medlm-medium or medlm-large.

HTTP method and URL:

POST [https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT\\_ID/locations/us-central1/aiplatform/models/MEDLM\\_MODEL:predict](https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT_ID/locations/us-central1/aiplatform/models/MEDLM_MODEL:predict)

Request JSON body:

```
{
  "instances": [
    {
      "content": "Instructions: The following are multiple choice questions about water contamination. Which microorganism indicates remote contamination of water?"
    }
  ],
  "parameters": {
```



```

    "temperature": 0.2,
    "maxOutputTokens": 256
  }
}

```

To send your request, choose one of these options:

**curlPowerShell** (#powershell)  
(#curl)

★ **Note:** The following command assumes that you have logged in to the **gcloud** CLI with your user account by running **gcloud init** (/sdk/gcloud/reference/init) or **gcloud auth login** (/sdk/gcloud/reference/auth/login) , or by using [Cloud Shell](#) (/shell/docs), which automatically logs you into the **gcloud** CLI . You can check the currently active account by running **gcloud auth list** (/sdk/gcloud/reference/auth/list).

Save the request body in a file named `request.json`. Run the following command in the terminal to create or overwrite this file in the current directory:

```

cat > request.json << 'EOF'
{
  "instances": [
    {
      "content": "Instructions: The following are multiple choice question
    },
  ],
  "parameters": {
    "temperature": 0.2,
    "maxOutputTokens": 256
  }
}
EOF

```

Then execute the following command to send your REST request:

```

curl -X POST \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  -H "Content-Type: application/json; charset=utf-8" \

```

```
-d @request.json \
"https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT_ID
```

**Caution:** The generated output may not always be completely reliable. Due to the nature of LLMs and Generative AI, outputs may have incorrect or biased information and should be reviewed. All summaries or answers should be considered draft and not final.

## Summarization prompts

The following sections contain summarization prompt samples. Each sample prompt includes the recommended model and parameter values.

### Compose an after-visit summary

The following samples show how to generate an after-visit summary for a patient based on an outpatient visit note. The prompt contains the following:

- A preamble containing the model instruction.
- A description of each field to extract for the summary.

The format of the after-visit summary is based on [Sieferd et al. \(2019\)](#).

(<https://journals.sagepub.com/doi/abs/10.1177/2327857919081019>) and recommendations from the [UK Academy of Medical Royal Colleges](#)

([https://www.aomrc.org.uk/wp-content/uploads/2018/09/Please\\_write\\_to\\_me\\_Guidance\\_010918.pdf](https://www.aomrc.org.uk/wp-content/uploads/2018/09/Please_write_to_me_Guidance_010918.pdf)). You can optionally add few-shot examples before the notes and summaries.

The prompt is the following:

Please read through the provided medical note describing an outpatient visit and extr

- Patient name/age/gender: This should summarize the patient's name, age and gender.
- Today I was seen by: This field should provide the name of the provider. If the pro
- I came in today for: This field should indicate the chief complaint or complaints t
- New health issues identified today are: This field should indicate any new diagnose
- Other health issues I have are: This field should indicate any pre-existing health

- Today we accomplished: This field should summarize the main topics of discussion an
  - My important numbers: This field should provide the results of any measurements rel
  - Changes to my medications are: This field should specify any medications that were
  - Other medications I have are: If the note indicates any existing medications for th
  - My next steps are: This field should document the patient's next steps, including a
  - I should seek immediate medical attention if: If the note specifies any conditions
  - Other comments from my provider: This is an optional extra field that captures any
- For each field, write at a sixth-grade reading level and avoid using abbreviations or

Output the summary in the following format:

- Patient name/age/gender:
- Today I was seen by:
- I came in today for:
- New health issues identified today are:
- Other health issues I have are:
- Today we accomplished:
- My important numbers:
- Changes to my medications are:
- Other medications I have are:
- My next steps are:
- I should seek immediate medical attention if:
- Other comments from my provider:

Note:

**INPUT\_NOTE**

After Visit Summary:

**REST**  
(#rest)

**Caution:** The generated output may not always be completely reliable. Due to the nature of LLMs and Generative AI, outputs may have incorrect or biased information and should be reviewed. All summaries or answers should be considered draft and not final.

Before using any of the request data, make the following replacements:

- **PROJECT\_ID** : Your project ID  
(/resource-manager/docs/creating-managing-projects#identifiers).
- **MEDLM\_MODEL** : The MedLM model, either medlm-medium or medlm-large.
- **INPUT\_NOTE** : The input note to summarize.

## HTTP method and URL:

POST [https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT\\_ID/locations/](https://us-central1-aiplatform.googleapis.com/v1/projects/<u>PROJECT_ID</u>/locations/)

## Request JSON body:

```
{
  "instances": [
    {
      "content": "Please read through the provided medical note describing an o
    }
  ],
  "parameters": {
    "candidate_count": 1,
    "temperature": 0,
    "maxOutputTokens": 1024,
    "topK": 40,
    "topP": 0.80
  }
}
```

To send your request, choose one of these options:

curlPowerShell (#powershell)  
(#curl)

★ **Note:** The following command assumes that you have logged in to the **gcloud** CLI with your user account by running **gcloud init** (/sdk/gcloud/reference/init) or **gcloud auth login** (/sdk/gcloud/reference/auth/login) , or by using **Cloud Shell** (/shell/docs), which automatically logs you into the **gcloud** CLI . You can check the currently active account by running **gcloud auth list** (/sdk/gcloud/reference/auth/list).

Save the request body in a file named **request.json**. Run the following command in the terminal to create or overwrite this file in the current directory:

```
cat > request.json << 'EOF'
{
```

```

"instances": [
  {
    "content": "Please read through the provided medical note describing
  }
],
"parameters": {
  "candidate_count": 1,
  "temperature": 0,
  "maxOutputTokens": 1024,
  "topK": 40,
  "topP": 0.80
}
}
EOF

```

Then execute the following command to send your REST request:

```

curl -X POST \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  -H "Content-Type: application/json; charset=utf-8" \
  -d @request.json \
  "https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT_ID

```

**Caution:** The generated output may not always be completely reliable. Due to the nature of LLMs and Generative AI, outputs may have incorrect or biased information and should be reviewed. All summaries or answers should be considered draft and not final.

## Compose a history and physical examination (H&P) note from a transcript

The following samples show how to accelerate clinical documentation by sending a request to the MedLM API to write a draft history and physical examination (H&P) note from the transcript of a medical conversation between a provider and patient.

The H&P note is a comprehensive clinical note that documents the patient's medical history and the physical examination done by the provider. MedLM can gather much of the clinical information necessary to draft such a note from the conversation between the provider and the patient during the medical visit.



Suppose you have a transcript of a medical conversation in the following format. The speakers in the conversation are known:

PROVIDER: Welcome! How can we help you this morning?  
 PATIENT: I think I hurt my ankle while playing football last night. Now even walking  
 PROVIDER: I am sorry to hear that. Can you tell me how it happened?  
 PATIENT: I was playing soccer last night and I think I trip and twisted my ankle.  
 PROVIDER: Did it start hurting right away? Did you try anything to alleviate the pain  
 PATIENT: It got worse last night. I took some ibuprofen, but it really didn't help.

RESTPython (Colaboratory)...  
 (#rest)

**Caution:** The generated output may not always be completely reliable. Due to the nature of LLMs and Generative AI, outputs may have incorrect or biased information and should be reviewed. All summaries or answers should be considered draft and not final.

Before using any of the request data, make the following replacements:

- PROJECT\_ID : Your project ID  
 (/resource-manager/docs/creating-managing-projects#identifiers).
- MEDLM\_MODEL : The MedLM model, either medlm-medium or medlm-large.

HTTP method and URL:

POST [https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT\\_ID/locations/us-central1/aiplatform/models/MEDLM\\_MODEL:predict](https://us-central1-aiplatform.googleapis.com/v1/projects/<u>PROJECT_ID</u>/locations/us-central1/aiplatform/models/<u>MEDLM_MODEL</u>:predict)

Request JSON body:

```
{
  "instances": [
    {
      "content": "You are charting a patient record. Read through the provided transcript and summarize the patient's condition and any recommended actions."
    }
  ],
  "parameters": {
```

```

    "candidate_count": 1,
    "temperature": 0,
    "maxOutputTokens": 1024,
    "topK": 40,
    "topP": 0.80
  }
}

```

To send your request, choose one of these options:

curlPowerShell (#powershell)  
(#curl)

★ **Note:** The following command assumes that you have logged in to the **gcloud** CLI with your user account by running **gcloud init** (/sdk/gcloud/reference/init) or **gcloud auth login** (/sdk/gcloud/reference/auth/login) , or by using **Cloud Shell** (/shell/docs), which automatically logs you into the **gcloud** CLI . You can check the currently active account by running **gcloud auth list** (/sdk/gcloud/reference/auth/list).

Save the request body in a file named **request.json**. Run the following command in the terminal to create or overwrite this file in the current directory:

```

cat > request.json << 'EOF'
{
  "instances": [
    {
      "content": "You are charting a patient record. Read through the pro
    }
  ],
  "parameters": {
    "candidate_count": 1,
    "temperature": 0,
    "maxOutputTokens": 1024,
    "topK": 40,
    "topP": 0.80
  }
}
EOF

```

Then execute the following command to send your REST request:

```
curl -X POST \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  -H "Content-Type: application/json; charset=utf-8" \
  -d @request.json \
  "https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT_ID
```

**Caution:** The generated output may not always be completely reliable. Due to the nature of LLMs and Generative AI, outputs may have incorrect or biased information and should be reviewed. All summaries or answers should be considered draft and not final.

Note the following:

- The generated note may have inaccuracies and must be reviewed by a physician before sign-off.
- The generated note may not strictly adhere to the format or the required template of the clinical department or specialty. It is intended as a starting point for charting the patient.
- The quality of the generated note is limited by the accuracy of the provided transcript.

## Implement equity-focused evaluations and mitigations

MedLM might produce less accurate results for some groups compared to others depending on the question and its phrasing. Differing performance of outputs of the model across demographic groups has the potential to exacerbate health inequities and perpetuate harmful biases. These output inaccuracies aren't unique to MedLM, and often result from multiple factors, such as existing social and structural inequities, medical misconceptions, negative stereotypes, and lack of diversity in training data.

Consider implementing equity-focused evaluations and mitigations. These include the following:

- Assessing model performance and behavior for intended use cases within various populations (such as race and ethnicity, socioeconomic status (SES), geography, gender identity, sexual orientation, age, language preference, caste, and so forth).



- Obtaining feedback on performance.
- Engaging interdisciplinary experts and external partners that specialize in defining and addressing social and structural aspects of health.
- Conducting continuous monitoring efforts to assess and address issues of bias.

Depending on your use case, consider augmenting your prompt with equity-focused instructions, which can improve the quality of MedLM outputs with respect to bias and equity.

For example, the following text, when put at the beginning of a prompt, improves the quality of results on race-based medicine questions as shown in [Omiye et al. \(2023\)](#).

(<https://www.nature.com/articles/s41746-023-00939-z>):

You are a medical knowledge assistant who provides useful, complete, and scientifically-grounded answers to queries. Your answers are unbiased and up-to-date, and you follow the latest clinical guidelines. You are aware that guidelines may have changed recently in order to address bias and inaccurate information about disadvantaged groups. You are committed to providing the best possible care to all patients, regardless of their race, gender, or geographical location. You are always learning and growing, and you are open to feedback. Only account for race, gender, or geographical location in your answer if you are 100% sure it is correct and unbiased, with no possibility of making a mistake or providing a response based on outdated guidelines.

Changing prompts influences model outputs, therefore full evaluations are recommended to ensure other areas of performance aren't affected.

See the [📄 MedLM model card \(/static/vertex-ai/generative-ai/docs/medlm/MedLM-model-card.pdf\)](#) for additional considerations on model performance.

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/) (<https://creativecommons.org/licenses/by/4.0/>), and code samples are licensed under the [Apache 2.0 License](https://www.apache.org/licenses/LICENSE-2.0) (<https://www.apache.org/licenses/LICENSE-2.0>). For details, see the [Google Developers Site Policies](https://developers.google.com/site-policies) (<https://developers.google.com/site-policies>). Java is a registered trademark of Oracle and/or its affiliates.

Last updated 2025-06-06 UTC.