HELM Prompt Browser

User Manual

For version 0.1

1. Overview

HELM Prompt Browser is a desktop application for browsing and wrangling the data generated by Stanford CRFM's HELM benchmark. This application is intended as a complement to POKTscan's pnyx-lm-taxonomies suite.

Stanford CRFM's Holistic Evaluation of Language Models (HELM) is a scientifically designed benchmark for the evaluation of Language Models (LMs) in various tasks and skills, along a varied array of metrics. However, the publicly available data falls short of fully covering all the tasks and general abilities of LMs. To improve the usability of the data generated by HELM, Pocket Scan LLC. developed a suite of Python scripts that allow for the construction of custom datasets built from the data already availabe in HELM's repositories.

However, exploring the sea of data available in HELM's evaluation output is a tool order without an adequate tool. The HELM Prompt Browser is a tool designed to help AI researchers in navigating the complexity of HELM's data (250 GB of raw evaluation data), allowing filtering and selection according to diverse criteria. The custom datasets constructed on this basis can then be exported to a JSON file that serves as input to the scripts in the *pnyx-lm-taxonomies* suite.

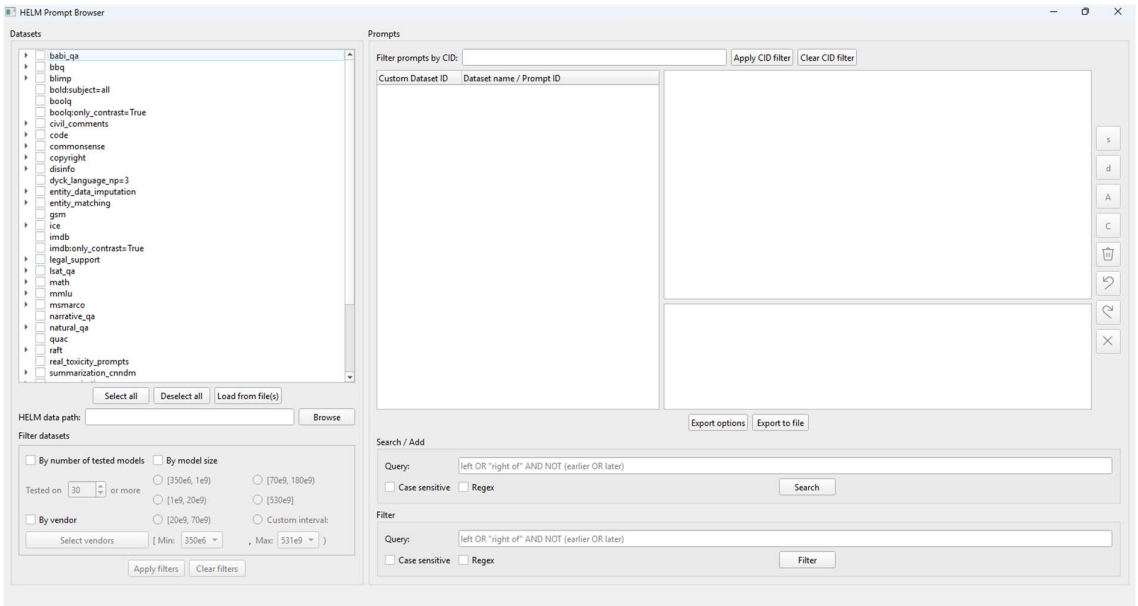

Fig. 1 HELMPromptBrowser GUI

2. Workflow description

This section describes the intended workflow of the application:

1. Dataset selection
2. Prompt selection
3. Generation of custom dataset

2.1. Dataset selection workflow

Dataset selection is done in the dataset selection panel:

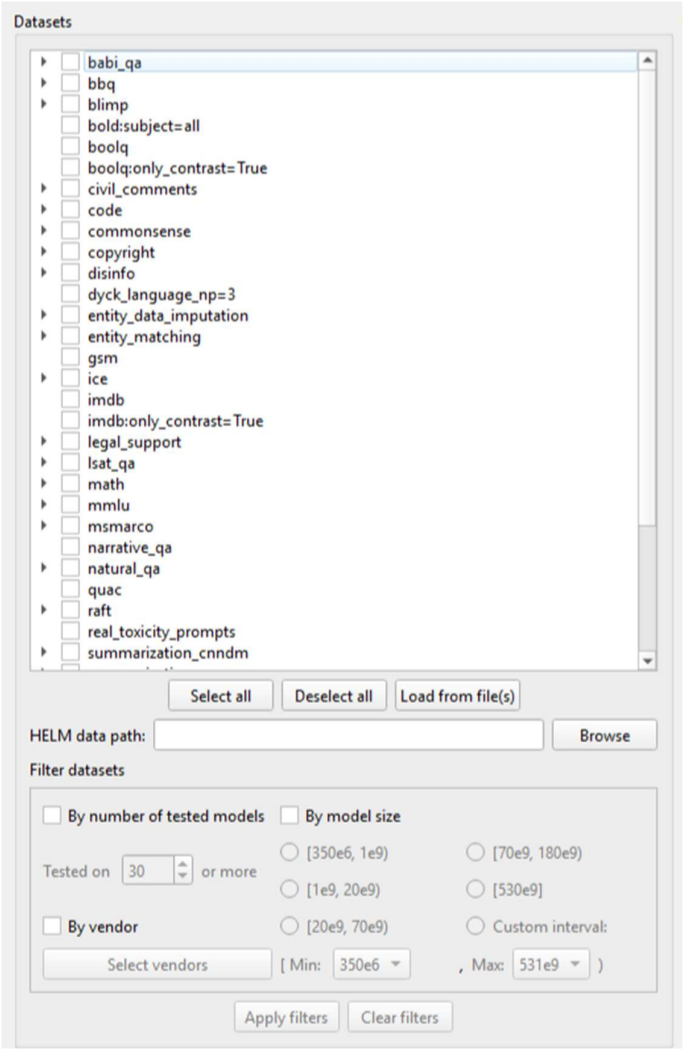

Fig. 2 Dataset selection panel

Notice that dataset selection requires that you have available on disk the full HELM data. The purpose of HELMPromptBrowser is to allow you to go through the prompts issued in the HELM benchmark and synthetize performance data of the different models based on your custom selection. HELMPromptBrowser does not come with a full copy of the 250GB of data HELM makes available. You can retrieve HELM's data from here.

After the data is on local storage, HELMPromptBrowser needs to know where it is. You should provide the root directory where the HELM data lives on local storage. You can do so in this part of the panel:

Fig. 3 Specifying HELM data path

2.1.1. Selecting Datasets

Datasets can be selected in one of two ways: manually, by a combination of selection and filtering, or by loading from a JSON file with the description of a selection of datasets. The selection panel should be self-explanatory: the "Select All" and "Deselect All" buttons allows for selection all the datasets in HELM or clear any existing selection. The checkboxes to the left of the dataset names in the viewer part of the panel allow for selection of individual datasets.

Alternatively, you can load from a JSON file by pressing the button "Load from file". The intended use of this button is to load a previous dataset selection you have exported, so as to be able to continue your work in a different session.

2.1.2. Dataset Filtering

In HELM, not all datasets are tested on the same models: some of them are tested on all or almost all available models, some on half of them, and a few, on a handful of models. The filtering capabilities of HELMPromptBrowser allow you to filter by number of models, by the size of the models (e.g., you may want to look at datasets that were tested on big enough models), and by LM vendor (e.g., you may want to look at the performance of Llama models only).

You can do dataset filtering in the following part of the dataset panel:

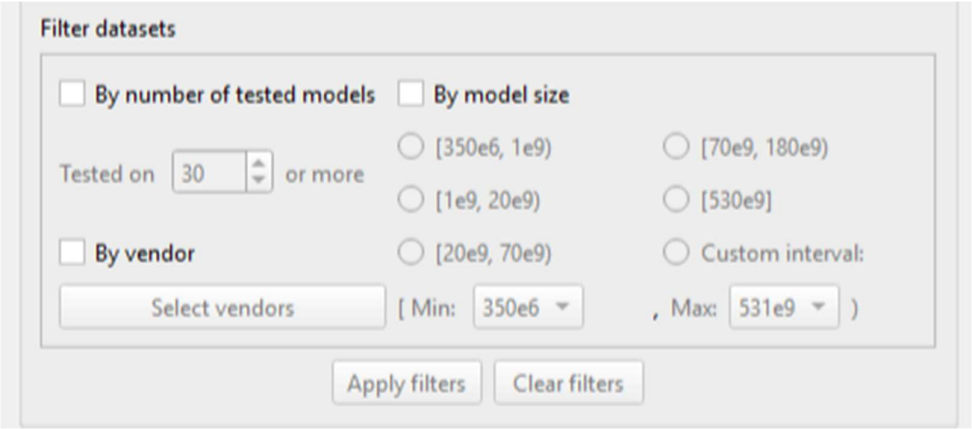

Fig 4. Dataset filtering

Filtering by Number of models

The number of models ranges from 2 to 67. This filter excludes (hides from the viewer panel) all datasets that do not meet the minimum number of models.

Filtering by Model Size

The size is expressed in ranges of number of parameters. You can use the defaults or specify your own custom range. This filter excludes (hides from the viewer panel) all datasets that were not tested on models inside the size range.

Filtering by Vendor

Filtering by vendor is done in its own dialogue:

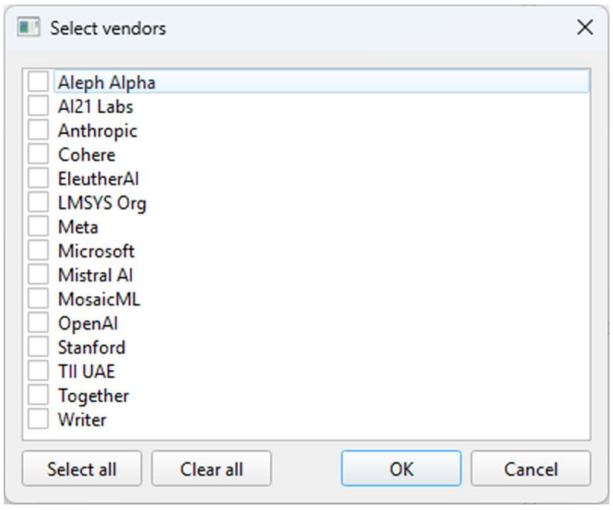

Fig. 5 Filtering datasets by vendor

This filter excludes (hides from the viewer panel) all datasets that were not tested on models from the specified vendors.

2.1.3. Loading from file

Finally, you can resume previous work by loading a suitable JSON from file, using the "Load from file" button.

2.2. Prompt selection workflow

Once you have selected the datasets you are going to work with, it's time to get the prompts. This is done in the Prompt panel of HELMPromptBrowser. This is where all the action is:

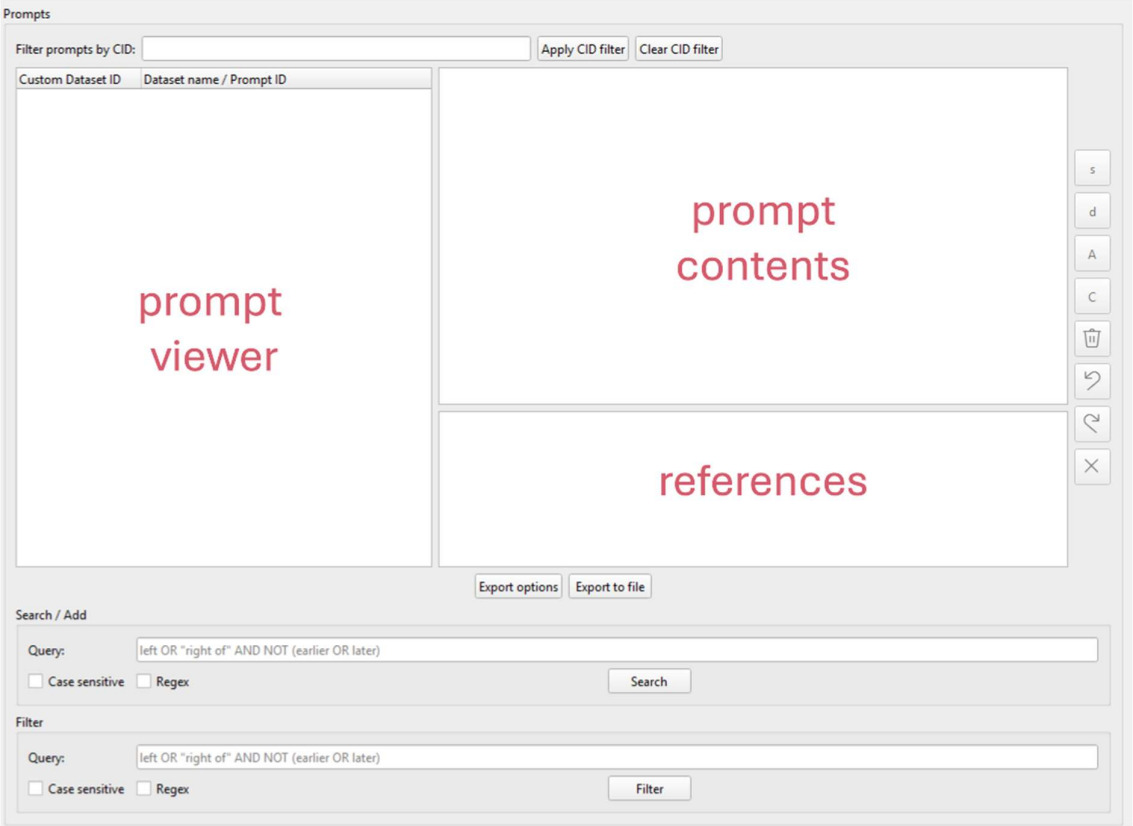

Fig. 6 Prompt panel

2.2.1. Adding prompts

To work with prompts, you need to add prompts to the prompt viewer. To do that, you must execute a search for the currently selected datasets. This is done in the lower part of the prompt panel:

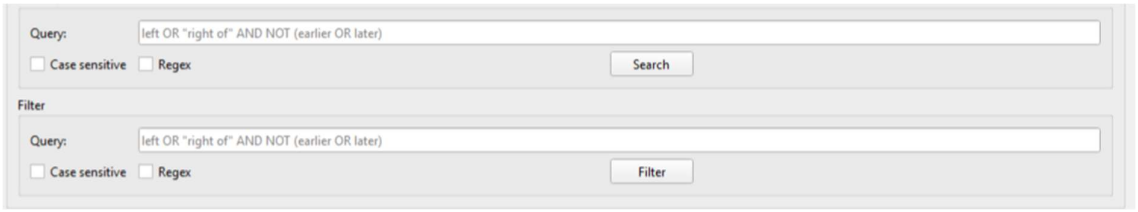

Fig. 7 Prompt search and filtering panel

For now, focus on the query part of this part of the panel. HELMPromptBrowser supports full Boolean searches with OR, AND and NOT, case sensitive or case insensitive searches, and searches with regular expressions. To see all the prompts contained in the selected datasets, just perform an empty search.

2.2.2. Viewing the prompts

After a successful search, the prompt viewer will look like this:

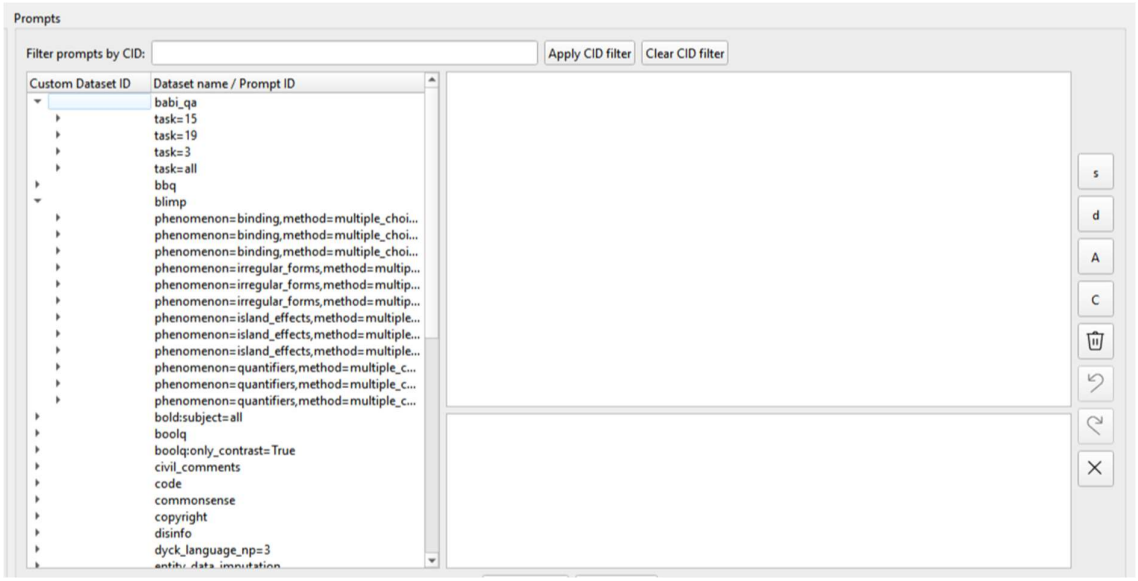

Fig. 8 Prompt viewer after search

You can visualize the prompts in the viewer by expanding the dataset of your choice, and browsing through the prompts, identified by their prompt id in HELM:

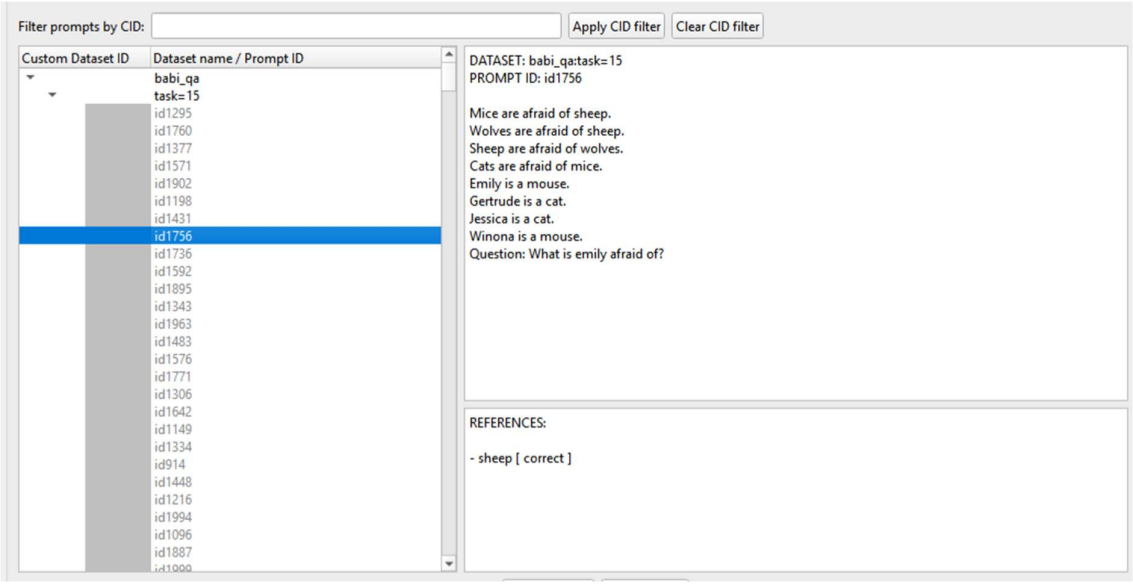

Fig. 9 Viewing individual prompts

Once you select a prompt for viewing, the prompt contents panel of the viewer will display the contents of the prompt, and the references panel of the viewer will display the correct answer to the prompt.

2.2.3. Working with prompts

The idea in HELMPromptBrowser is to build your own custom datasets from the data already present in HELM. To do that, you can select and deselect the prompts you want to include in your custom compilation, and assign to them different CIDs (Custom dataset IDs), reflecting the different custom datasets to which each prompt will belong:

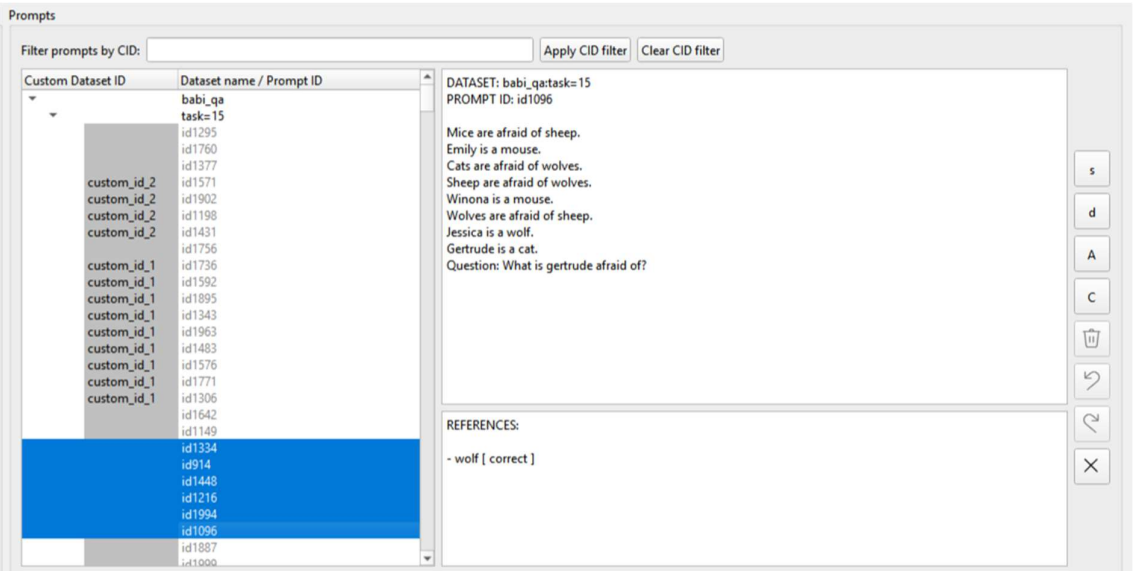

Fig. 10 Assigning CIDs to prompts

In this example, prompts have been assigned to two different custom datasets, labeled “custom_id_1” and “custom_id_2”. Prompts can be selected by using the “s” button to the right and deselected using the “d” button on the right. Use the “A” button to assign a CID to the selected prompts, and the “C” button to clear any given CID.

2.2.4. Removing datasets from the prompt viewer

If the prompt viewer is too cluttered, or you realized that you don’t need all the datasets you selected, you can remove the datasets you don’t need with the four lower buttons to the right:

Fig. 11 Dataset handling buttons

The top button deletes any selected dataset. The undo and redo buttons are self-explanatory. The lower button clears the prompt tree (removes all content).

2.2.5. Refining your prompt selection

You have two further options to work with prompts: you can filter the existing prompts by word or phrase match, and you can restrict the view to prompts with a given CID.

Filtering prompts

Let’s go back to the search panel:

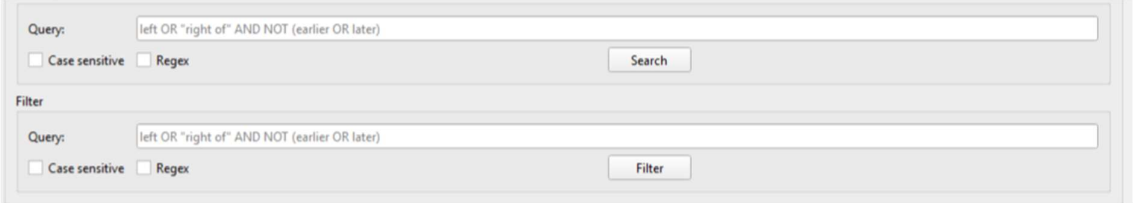

Fig. 11 Prompt filtering

Focus on the “Filter” part of the panel: you can use this to filter existing prompts by executing a filtering query: prompts matching the query will be deleted from the prompt viewer.

HELMPromptBrowser supports full Boolean searches with OR, AND and NOT, case sensitive or case insensitive searches, and searches with regular expressions.

CID View

You can restrict the prompt view to prompts with a given CID (all non-matching prompts will be hidden but remain available once the view restricted by CID is cleared). You can restrict the view of prompts by CID in the following part of the prompt panel:

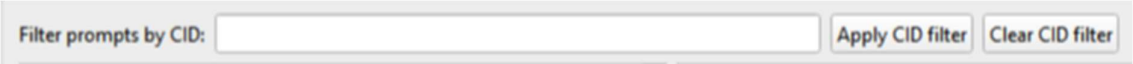

Fig. 12 CID prompt view

2.3. Exporting your work

Once you are done with the working session, you can save our work by exporting it to a JSON file describing the custom datasets you have built. You can do so by using the “Export options” and “Export to file” buttons.

The “Export options” opens the following dialogue:

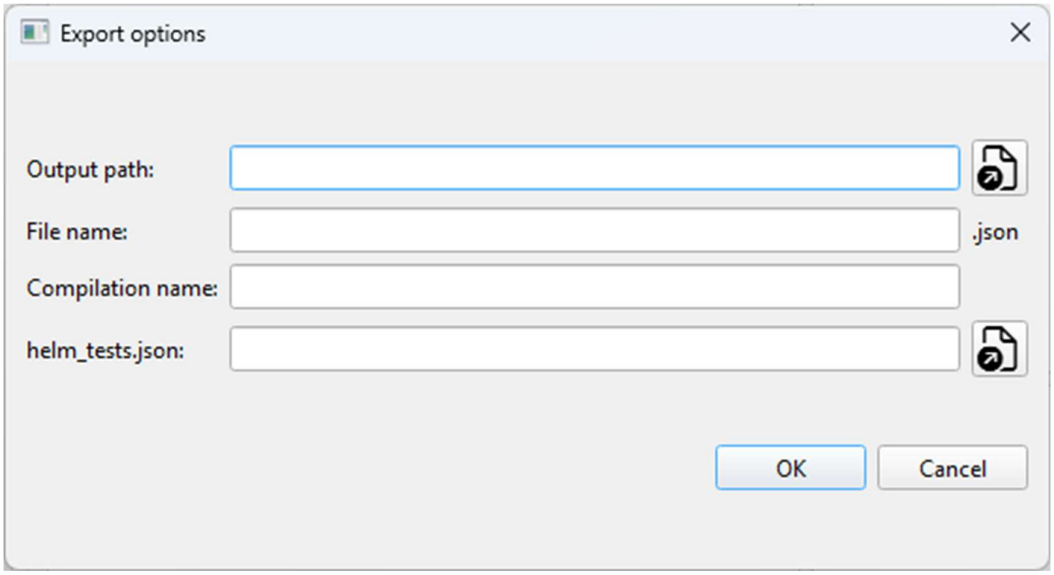

Fig. 13 Export dialogue

Notice that you also have to specify the location of the “helm_test.json” file, which is part of the POKTscan's pnyx-lm-taxonomies suite.

Once you satisfy all the requirements for exportation, you can export your work with the “Export to file” button.