

NIMLAS Workshop 2026 Manual: Evaluating Open-Source Large Language Models for Research

Why Do We Need an Evaluation Framework?

New open-source language models are released almost every month.

Some are faster. Some are larger. Some are better at reasoning. Some are better at coding.

Others specialize in multilingual text.

Rather than asking "**Which model is the best?**", researchers should ask "**Which model is best for my research project?**"

This guide provides a practical framework for answering that question.

Step 1. Define Your Research Task

Before evaluating any model, clearly identify what you want the model to do.

Checklist

- ☐ Summarize interviews
- ☐ Code qualitative responses
- ☐ Identify themes
- ☐ Extract structured information
- ☐ Draft reports
- ☐ Explain statistical results
- ☐ Translate documents
- ☐ Other: _____

Step 2. Understand Your Data

Checklist

What kind of data will you analyze?

- ☐ Interview transcripts
- ☐ Focus groups
- ☐ Survey responses
- ☐ News articles
- ☐ Policy documents
- ☐ Social media
- ☐ PDFs
- ☐ Mixed text sources

Data Size

Approximately

- ☐ Less than 100 documents
- ☐ 100–1,000 documents
- ☐ More than 10,000 documents

Sensitive Information?

- ☐ No
- ☐ Yes

If yes,

consider running the model locally.

Step 3. Evaluate Candidate Models

Use the following checklist for every model you consider.

Table 1: Worksheet for Choosing Candidate Models

Criterion	Excellent	Good	Fair	Poor
Produces accurate summaries	☐	☐	☐	☐
Correctly extracts information	☐	☐	☐	☐
Hallucinates rarely	☐	☐	☐	☐
Handles long documents well	☐	☐	☐	☐
Easy to install	☐	☐	☐	☐
Runs smoothly on my computer	☐	☐	☐	☐
Produces answers quickly	☐	☐	☐	☐
Documentation is clear	☐	☐	☐	☐
Active developer community	☐	☐	☐	☐

Step 4. Test the Model Yourself

Never rely only on benchmark scores. Instead, test the model using **your own research tasks**.

Test 1

Summarize an interview.

Evaluation

☐ Accurate

☐ Concise

☐ No hallucinations

☐ Easy to read

Test 2

Extract variables.

Evaluation

- ☐ Correct variables
- ☐ No missing information
- ☐ Structured output
- ☐ Easy to verify

Test 3

Identify themes.

Evaluation

- ☐ Themes are meaningful
- ☐ Themes are supported by evidence
- ☐ Easy to interpret

Step 5. Evaluate Reliability

Ask the model the same question multiple times.

Checklist

- ☐ Similar answers each time
- ☐ Different wording but same meaning
- ☐ No contradictory conclusions
- ☐ Stable classifications

Large differences may indicate the need for better prompts or another model.

Step 6. Evaluate Hallucinations

Ask questions whose answers can be verified easily.

Example

Interview: I live in Michigan.

Question: What is the participant's occupation?

Expected answer: Not mentioned.

Evaluation

☐ Model correctly answered

"Not mentioned"

OR

☐ Model invented an occupation

Models that frequently invent information may require more careful prompting.

Step 7. Evaluate Prompt Sensitivity

Use three different prompts for the same task.

Example

Prompt A: Summarize.

Prompt B: Summarize in three sentences.

Prompt C: Summarize in three sentences using only evidence from the interview.

Questions

Did the answers improve?

Did hallucinations decrease?

Did the format become more useful?

Models that respond well to improved prompts are often easier to work with.

Step 8. Evaluate Practical Considerations

Table 2: Questions for Evaluating Models

Question	Yes	No
Runs on my laptop	☐	☐
Response time is acceptable	☐	☐
Easy to install	☐	☐
Fits available storage	☐	☐
Works offline	☐	☐
License allows my intended use	☐	☐

Sometimes the fastest model is the most useful model.

Step 9. Compare Candidate Models

Table 3: Comparing Candidate Models

Criterion	Model A	Model B	Model C
Summarization			
Theme generation			
Information extraction			
Hallucination rate			
Speed			
Ease of installation			
Hardware requirements			
Overall impression			

Do not compare only one task. Evaluate the tasks that matter for your research.

Step 10. Make Your Decision

Use this checklist.

- ☐ Accurate enough
- ☐ Reliable
- ☐ Easy to use
- ☐ Fits my computer
- ☐ Meets privacy requirements
- ☐ Appropriate license
- ☐ Works well for my research task

If most boxes are checked, the model is probably a good candidate.

A Researcher's Evaluation Rubric

Score each category from 1 (Poor) to 5 (Excellent).

Table 4: Model Evaluation Scoreboard

Category	Score
Accuracy	
Hallucination resistance	
Prompt responsiveness	
Ease of use	
Speed	
Privacy	
Documentation	
Overall usefulness	

Total Score: ____ / 40

Remember: The highest-scoring model is not necessarily the best model. Choose the model that best supports your own research.

Final Checklist Before Starting a Research Project

Research

- ☐ Research question defined
- ☐ Appropriate prompts prepared
- ☐ Evaluation plan developed

Model

- ☐ Model tested
- ☐ Model documented
- ☐ Version recorded

Reproducibility

- ☐ Prompt saved
- ☐ Model name recorded
- ☐ Model version recorded
- ☐ Output verified

Ethics

- ☐ Privacy protected
- ☐ Sensitive data handled appropriately
- ☐ Human review completed

Takeaway: Open-source LLMs will continue to evolve. Your ability to evaluate them critically is more valuable than memorizing the name of today's most popular model. The best researchers are not those who always use the newest model. They are the ones who can thoughtfully select, evaluate, document, and justify the model they use.