

## Supplemental Section 2: Blueprint for Using LLM Cluster Naming on Your Project

### Is It Appropriate?

When considering an LLM cluster labeling process you should first determine whether the approach is appropriate for the project and the context that the project is operating in. Although interest in LLM and GenAI applications is growing rapidly, the acceptance of such methods varies widely across federal and nongovernmental agencies and among researchers. Discussions with the project team should be centered around ensuring that the approach can achieve an accurate, trustable, and cost-effective solution. Some sample questions and areas of focus could include the following:

- **Stakeholder Engagement:** Both the clients and the internal project leadership should be willing and interested to engage in this type of solution. Because this is a new approach, the project team should be prepared to help even enthusiastic clients with information about how the approach works and why the results can be received with confidence.
- **Tolerance for Error:** Based on the type of data and the project (e.g., restaurant reviews vs. 911 calls), what is the client's tolerance for "error" in the output? As the tolerance for error decreases the need for human review will increase and the cost savings of an LLM approach will diminish.
- **Cost Effectiveness:** Although the cost factors for a traditional labeling approach are straightforward to determine, estimating the cost for an LLM cluster labeling approach will involve determining a level of effort for the programming tasks, human reviewers and adjudicators, and a potentially variable number of development cycles as the project team builds confidence in the results and refines their methods. Additional cost considerations might be how often the client expects to perform this task and whether any economies of scale may influence the decision. The larger the number of clusters, the higher the traditional method's labor costs will be and the greater the potential cost savings of an automated approach.
- **Data Restrictions:** Restrictions on the use of the data may come from the data classification identified by the client, your organization's internal security and compliance team, data use agreements, or federal or state law. Client agreements may also impose limitations on the transfer of project data to a third party (e.g., external LLM service) regardless of any data classification. If the data being used are not clearly in the public domain, the project team needs to be able to identify and evaluate any such restrictions prior to submitting the data to an LLM.
- **The Big Picture:** Pretend that your cluster labeling has gone exactly as planned and imagine what the final output would be. Compare this output to what a traditional labeling process could generate and make sure that the approach still makes sense.

Many external factors such as the heterogeneity of the data, cluster quality, and document domain to LLM training data interactions can play a significant role in the result of a cluster labeling effort.

### Framing the LLM Cluster Labeling Process

Once a project has decided to move forward with an LLM cluster labeling approach the team can begin to identify the methods and reviews that they want to implement. A method for automated labeling can be thought of as a combination of selected input(s), prompt template(s), and feedback processes that may act on the LLM output (including resubmitting some form of the output to the LLM). In most cases it will not be practical to submit the entire document text for LLM evaluation because of current context windows so whatever input selection is made should be a reasonable representation of the underlying document or subset of documents. In this study, we decided that the prompt inputs would be keywords extracted from identified clusters, keywords drawn from sampled documents, and direct document attributes or fragments such as the document title.

Strategies for developing prompts and prompt templates are outside the scope of this guide but users are encouraged to explore different prompting approaches based on their personal knowledge of the data and current prompting research in the data's domain. A simplified prompt template might look something like "The following text samples are drawn from the XXX of a set of source documents in the field of XXX. Generate a descriptive label that identifies a cluster that all of these samples could belong to. Be as specific as possible."

With a series of prompts and inputs selected, the last step in determining the project's approach is the decision of what, if any, feedback processes will run on the LLM output. The simplest option would be to either take no action on the initial LLM output, or to provide some minor formatting for later usage. More involved and possibly more successful post-processing could include things like prompting the LLM to rate the provided cluster label(s) and provide a confidence rating, generating multiple rounds of labels from the same sampled source and prompting for a consensus label, or even setting up an Agent-Critic approach where two LLM instances are used with output from A being critiqued by C and then regenerated from A with that response. Although our research showed that across two data sources there was always an approach that was higher rated than human-generated labeling, there is no basis other than experience for having an a priori expectation that one method would perform better than another, or better than a human reviewer. Experimentation is key.

### Practical Considerations for Running Your Cluster Naming Pipeline

Executing your defined combinations of inputs, prompts, and feedback steps will likely be one of the easier parts of the implementation depending on how complicated your feedback processes may be. The LLM used in this study provides an API interface, which means that the consuming application you write can be in the language of your choice. For our work we used python with langchain to help integrate with the OpenAI API but the tool selection can be influenced by the skillsets of the team. Implementation options will vary by the vendor and model selected but some common issues to consider include the following:

API key management and billing are also important considerations. While different vendors will have their own approaches to key and account management, project teams need to proactively plan for how to manage their LLM access. Before you begin using an LLM, through an API or otherwise, make sure you can answer: How are we securing the API keys or credentials for this project? Can access credentials be shared between users or projects? How do we separate billing charges between projects? How are we documenting the prompting and responses to support reproducibility?

LLM response variability (temperature) is going to be an interesting aspect of assessing reproducibility of research for LLM-based projects. Unless there is a compelling reason to do otherwise, we should keep temperature to 0. Keep in mind that even a temperature of 0 does not make the LLM response deterministic, and you will see some level of variability between LLM runs even with the same inputs.

### Evaluating LLM Output

Cluster labeling tasks are particularly well suited to LLMs because the main qualitative evaluation is whether it looks like the best option to a human in the loop. The evaluation process can be defined along a spectrum of complexity and implemented according to your project's budget, timeline, and risk tolerance. In its simplest form, the series of LLM responses and sample source documents from the cluster can be presented to a subject matter expert (SME) who could identify the best approach for implementation. As more complex approaches are desired, the team can introduce tournament-style prompt refinement, human labeling teams, domain specific qualitative metrics, etc.

A significant question once a performant LLM approach has been identified is how we build and maintain confidence in the model's responses. This confidence will likely be based on how well the model initially performs, how well subsequent rounds of testing do once an LLM approach has been implemented, and how often the LLM responses are checked to ensure that it is still performing as expected. The fluency of a modern LLM response can easily be misconstrued as providing a broad "reasonable analysis" of the underlying data and may cause clients to expect flexibility and adaptability that the model may not be able to provide. All LLM/GenAI approaches should plan and budget for a level of cyclical review that allows us to maintain confidence in model performance. It's also advisable to standardize and retain the responses from human evaluators so they can be available for some future retraining effort to bring the model back in line with expected performance.

### Summary

Implementing an LLM cluster-labeling method can represent a successful strategy for clients looking for cost-effective solutions. Project teams will need to iterate on developing appropriate prompt templates, inputs, and feedback mechanisms that make sense for the labeling task. Experimentation with these components of the LLM cluster labeling pipeline will be crucial in finding the right approach. Client SMEs responsible for evaluating the labeling results should be involved early in the process to provide insight to the project team and help build their own familiarity and confidence with the overall process.