

CEDAR: Context Engineering for Agentic Data Science

Rishiraj Saha Roy¹ Chris Hinze¹ Luzian Hahn¹ Fabian Küch¹

Abstract

We demonstrate CEDAR (Context Engineering for Data science with Agent R

1. Introduction

Motivation. In traditional data science (DS), a human expert, the data scientist, writes scripts for entire pipelines, including standard steps like data preprocessing, feature engineering, hyperparameter tuning, and finally computing metrics and visualizing insights. However, this work is tedious and repetitive, and can be significantly optimized with modern LLMs. For example, many users are exploring options like ChatGPT Advanced Data Analysis, where one simply uploads data and articulates entire problems as prompts.

However, such approaches have several limitations: (i) concrete instructions for real DS projects are more complex than can be solved via one-shot prompting; (ii)

capabilities of generative models are still limited with respect to arbitrary mathematical computations; (iii) data files are often very large and cannot simply be uploaded as attachments (ChatGPT Advanced Data Analysis has a limit of 512 MB per file, many Kaggle files are much larger); (iv) there are often privacy concerns and users may feel insecure uploading their enterprise data to cloud-based LLMs; and (v) DS solutions are typically multi-step: as a solution workflow progresses, simply packing all instructions, text, code, data, and results into the running context makes it unintelligible for most LLMs, and often exceeds context length limits.

Limitations of state-of-the-art. Agentic systems now power many real-world information retrieval (IR) and machine learning (ML) applications, where LLMs perform various roles in complex task pipelines (see, for example, (He et al., 2024), (Chu et al., 2025), (Qiu et al., 2024)). DS is a unique instance of such complexity, tightly coupling IR, ML, and traditional statistics. Agentic systems with effective context management can alleviate most of the above problems for DS tasks. This has become an active research area (Maojun et al., 2025; Jing et al., 2024), where the best systems usually achieve impressive results via sophisticated prompt chaining. But we found that it is very difficult for a typical end-user to get a clear idea into how the task is actually being solved (Grosnit et al., 2024; Guo et al., 2024; Hong et al., 2024). The closest work to ours’ is the very recent Jupyter Agent 2 (Colle et al., 2025) from Hugging Face. However, contrary to our application, Jupyter Agent 2 does not run locally: data needs to be uploaded to its cloud, and the time taken depends heavily on one’s network bandwidth (usually very slow). Also, the privacy of the uploaded data is not guaranteed. The purpose of our system CEDAR is to bring transparency and simplicity into DS solutions with LLMs.

Approach. CEDAR relies on effective *context engineering*, collectively referring to strategies for maintaining an optimal set of tokens during LLM inference, including all information that may land there outside of prompts¹. An overview of our workflow is in Fig. 1. The data scientist formulates the task as a structured prompt, that is passed on to an orchestrator agent. This agent routes text and code generation requests to sub-agents as tools, to generate a readable solution with

¹Department of Generative AI, Fraunhofer IIS, Am Wolfsmantel 33, 91058 Erlangen, Bavaria, Germany. Correspondence: rishiraj.saha.roy@iis.fraunhofer.de

¹<https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents>

short steps. The human inspects final outputs and revises instructions for further iterations, if necessary. Code will be released soon on GitHub after internal clearance.

Audience. The app can be used by beginner to expert-level IR, ML, NLP, and AI practitioners. IR and ML basics help, but there is no prerequisite of core DS knowledge to interpret the workflow: each solution step has natural language (NL) explanations. Beginners can get a feel of how basic data science tasks are solved. Intermediate users can contrast the faithfulness of the generated solution to the original intent, and gain insights into how agentic systems can be built for simplifying DS. Experts can scrutinize LLM-generated code snippets, and whether the generated solution mimics code written by a human data scientist.

Scope. Our system can solve beginner-level data science tasks from Kaggle. The current requirements are simply to have a clearly articulated goal and well-defined data.

2. Method

We now describe the key features of the CEDAR pipeline: prompts, the agentic setup, tool use, reliance on code for plan execution, and history rendering.

2.1. Structured prompts

To relieve the user from creating verbose prompts containing every pertinent task detail, we create a structured form that covers key aspects of most DS tasks. We split the project summary into high-level *general instructions* (estimated number of solution steps, expected number of plots, verbosity of plans, etc.) and *task-specific instructions* (with items like Task description, Data description, Data location, Metrics, Inputs, Outputs, and Special instructions). This makes DS tasks more understandable to LLMs.

2.2. Interleaved text and code

Instead of expecting the LLM to directly emit a final solution, say, like an optimized set of metrics, we make the system generate a human-readable workflow with succinct, numbered steps starting from data loading and all the way to model card generation. Each “*Step*” consists of a plan and a corresponding code snippet that implements this plan (Wang et al., 2024), like a Jupyter notebook. We then have an output with scrutable and reusable components like plans, code snippets, data snapshots, error traces, and intermediate plots.

2.3. Agentic setup

Such a workflow is enabled by the use of agents, i.e. different LLM instantiations with distinct prompts. We have three agents: a main orchestrator, and sub-agents for

text and code generation. The orchestrator decides whether to invoke the text/code generator (routing), or, seeing the blocks generated so far, that the *task is complete*. The calls to the sub-agents usually alternate between text and code like most Jupyter notebooks, but special situations may need consecutive text or code blocks. All agents receive the same context: the project summary, text and code snippets so far, and outputs of code snippet execution (see *history rendering* below for more details). They only vary in their responses: the main agent emits a sub-agent call, while the sub-agents emit text and code as per the current state of the workflow.

2.4. Tool calls

To prevent the orchestrator from generating free-form output that needs to be parsed, we implement the code and text generators as distinct *functions* so that they become *tools* to be called by the main agent². This has been a special feature in modern LLMs that highly facilitates building apps that chain LLM agents together. Further, to prevent hallucinated outputs like `write_code` when the correct function name is `request_code`, we use *structured outputs* by forcing the orchestrator to output JSON as per our schema³.

The communication with our tools is schema-driven, to prevent failures due to hallucinated responses by the main agent. The difference between “*spec*” (for `request_text`) and “*purpose*” (for `request_code`) is intentional, and it reflects the different goals of the two downstream agents (Text vs. Code). The Text Agent’s job is to produce Markdown explanations – human-readable commentary, analysis, or narration. The word “*spec*” (short for specification) captures what the Orchestrator wants the text to talk about – i.e., the topic or scope of the Markdown. For example:

```
{
  "action": "request_text",
  "spec": "Explain how the model will be evaluated
          and metrics that have to be computed."
}
```

The Text Agent writes a clear paragraph or bullet list about evaluation metrics – not code (we can think of “*spec*” as a creative or descriptive brief). The Code Agent must instead produce executable Python code. The word “*purpose*” indicates the goal or intent of that code cell — i.e., what this code is meant to achieve. For example:

```
{
  "action": "request_code",
  "purpose": "Load the training and test datasets
             into pandas DataFrames."
}
```

The Code Agent now generates the corresponding Python code to achieve this (“*purpose*” is a functional or operational

²<https://platform.openai.com/docs/guides/function-calling>

³<https://platform.openai.com/docs/guides/structured-outputs>

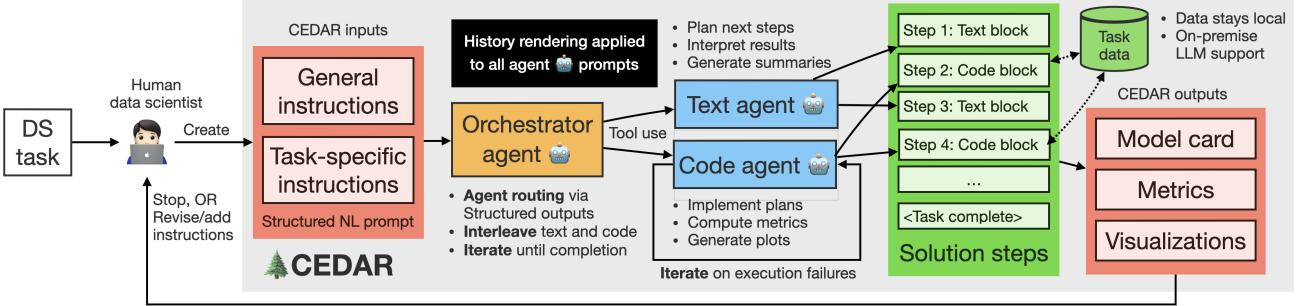


Figure 1: The CEDAR system with three LLM agents, facilitating a human data scientist’s job.

instruction). The difference helps both downstream agents interpret the Orchestrator’s intent unambiguously:

- the Text Agent focuses on narrative goals (explain, summarize, discuss); and
- the Code Agent focuses on technical goals (load, process, train, evaluate).

We could rename “*spec*” and “*purpose*” to a common field like “*instruction*”, but then both downstream roles would need logic to interpret that differently – which would make the orchestration less explicit and a bit more error-prone. The “*finish*” action is used when the Orchestrator decides that the notebook or analysis has reached a natural stopping point – the model has achieved its main goal and should wrap up. The accompanying `summary_hint` (optional) gives a short textual summary of what was achieved, key results, or next steps to note in the final cell. For example:

```
{
  "action": "finish",
  "purpose": "Baseline logistic regression trained and
             evaluated. Accuracy is about 0.72. No further
             steps required."
}
```

The finish response does not invoke any downstream LLM, but triggers the CEDAR app logic to create a final “*Finished*” cell in the notebook. Use of structured outputs reduces the amount of retries to solve specific sub-problems by separating routing from tool parametrization (Beurer-Kellner et al., 2024). A summary of orchestrator responses is in Table 1.

2.5. Code for math and data handling

Generating Python *code for computations* frees us from dependence on an LLMs’ math capabilities. Another significant advantage is that *data stays local*: the code operates on local data, and only snapshots and aggregate statistics are passed on to LLMs. We support on-premise LLMs, so that even such data digests need not leave the user’s system (in case working with sensitive data). We allow for *iterative code execution*: when the execution of a particular code block fails, the coder tool is prompted again with the current *error trace* and a request to accordingly *rewrite* the

code. This step adds robustness and enables recovery from name/type mismatches and missed imports, for instance.

Safety. Docker is used to execute code inside containerized runtime environments that isolate application processes from other host processes, thereby reducing potential security risks. Network access can be explicitly restricted through Docker’s networking configuration to prevent unintended data sharing or the execution of malicious external code. The execution environment is based on pre-built Docker images that include commonly used data science libraries and is extended as required.

2.6. History rendering

This module takes the task history, i.e., the list of all blocks so far (instructions, text, code, outputs, errors), as input, and converts them into a compact, LLM-friendly text summary as follows: (i) it appends user instructions; (ii) it numbers each text/code block (Text #3 or Code #5) for clarity and appends them in full as these are not very long (only adding outputs and not raw code loses vital rationale that led to these outputs). Actual *context bottlenecks* are code outputs and error messages, as handled next; (iii) for past code blocks, it adds only those that ran successfully; (iv) for these “*success*” blocks, it adds only *heads* of outputs as they carry the most vital information; (v) if the latest code block resulted in error, it adds the *tail* of the traceback instead, as it pinpoints the error most specifically; (vi) if the history so rendered exceeds 10k characters (configurable), only the most recent 10k are kept. This history is then passed on as context to the agent being invoked at any given moment.

3. Demonstration walkthrough

3.1. Using the app

Fig. 2 shows CEDAR solving a canonical Kaggle competition on LLM fine-tuning⁴. A user first selects

⁴<https://kaggle.com/competitions/llm-classification-finetuning>

Orchestrator output	Handled by	Meaning	Key argument	Output
request_text	Text Agent	Generate the next Markdown cell explaining results or outlining the plan for the next step	spec \mapsto Describes what to talk about (topic, focus, or goal of the text)	Explanatory Markdown block (added as a “text” cell)
request_code	Code Agent	Generate the next Python code cell that performs a concrete task (e.g. data load, training, evaluation)	purpose \mapsto Explains what the code should achieve	Executable Python code block (added as a “code” cell, executed immediately)
finish	CEDAR App Logic	Signal that the notebook is complete – no more steps are needed	summary_hint \mapsto Optional final note summarizing results	Marks the session as finished, and appends a final “finish” cell

Table 1: Explaining the use of Structured Outputs for possible orchestrator responses.

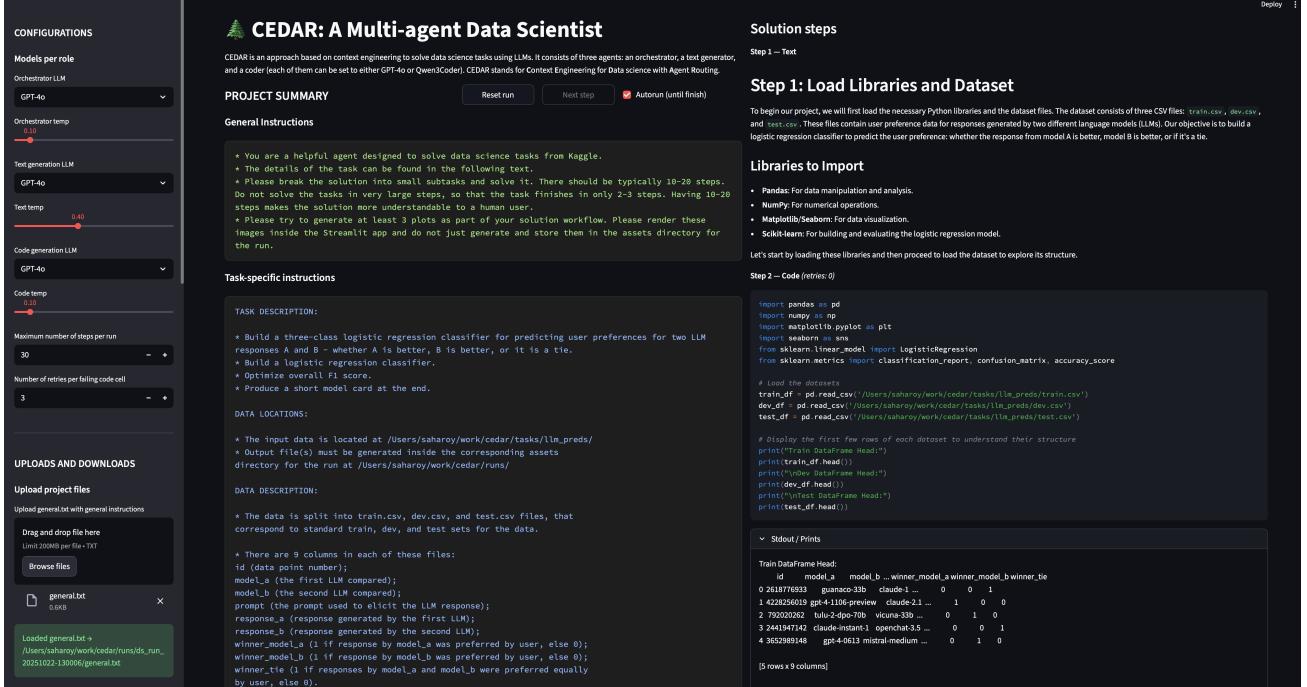


Figure 2: A screenshot of our CEDAR application, solving a Kaggle competition on predicting better LLM responses.

LLMs for orchestrator, text generator, and coder agents (GPT-4o via API, or Qwen3-Coder 30B⁵ locally via ollama⁶). Qwen3-Coder was chosen as an alternative as it was fine-tuned specifically for agentic coding. Next, general (green) and task-specific instructions (blue) need to be uploaded, instantly rendered as the “project summary” in the broad middle panel. The broad right panel is reserved for rendering the complete generated solution.

The user selects between “Next step” for inspecting a step at a time, and “Autorun” to see the full solution at once ($\simeq 3$ minutes for 10–20 steps). Steps can be “Reset” at any time.

3.2. Assets directory

An assets’ directory can be explored at any time during or after the run: it stores the structured prompt, generated

plots (also rendered online in the app), metrics, the model card and debug logs. We allow for *exporting* the solution as JSON, Markdown, or as a Jupyter notebook (making the solution editable), and *importing* a saved run as JSON. The run can be continued from where we left off when the saved JSON for a partially completed run is uploaded.

3.3. Configurations and debugging

The leftmost panel provides several knobs: (i) temperatures for each of our agents (to transition between response repeatability and diversity); (ii) maximum solution steps (default 30, suitable for most Kaggle tasks); and (iii) maximum code retries (default 3, solves commonly observed errors). An *autorun trace* shows running summaries for history rendering (the unpruned context is $\simeq 20k$ characters towards the end). If default tool calls fail, (often the case for Qwen3-Coder), we use *tool emulation*, forcing JSON outputs. Diagnostics for GPT-4o API keys and Qwen3 server

⁵<https://arxiv.org/abs/2505.09388>

⁶<https://ollama.com/library/qwen3-coder>

connections help resolve authentication and network issues, respectively. The context truncation limit (10k characters) and the head/tail size of standard output and standard error (20 lines) can also be set via code.

3.4. Backend and frontend

Our backend is in pure Python and does not use any special agent libraries. The frontend is built with Streamlit⁷. CEDAR can run on any laptop, as long as the RAM allows for loading the data in main memory. Qwen3-coder 30B is hosted on a GPU server (4x48GB NVIDIA Ada 6000 RTX, 512 GB RAM, 64 virtual cores).

4. Concluding remarks

Through our contribution, we show that building an agentic system pays off in the long run: the role of the human data scientist shifts from tediously scripting repetitive workflows to more cognitively rewarding tasks like structuring requirements, scrutinizing solutions, and suggesting optimizations.

Nevertheless, our focus here was on context engineering, and agent complexity in CEDAR is still rudimentary. Natural next steps would introduce independent agents that inspect a solution for faithfulness with respect to original user intent, and critique generated solution workflows for improving target metrics.

Acknowledgements

This work has been funded by the Free State of Bavaria in the DSGenAI project (Grant No.: RMF-SG20-3410-2-18-4). We thank members of the NLP team at Fraunhofer IIS for useful inputs at various stages of this work.

References

- Beurer-Kellner, L., Fischer, M., and Vechev, M. Guiding LLMs the right way: Fast, non-invasive constrained generation. *arXiv preprint arXiv:2403.06988*, 2024.
- Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., Zhong, A., Hu, X., Liang, J., Yu, P. S., et al. LLM agents for education: Advances and applications. *arXiv*, 2025.
- Colle, B., Yukhymenko, H., and von Werra, L. Jupyter Agents: training LLMs to reason with notebooks. <https://huggingface.co/spaces/lvverra/jupyter-agent-2>, 2025.
- Grosnit, A., Maraval, A., Doran, J., Paolo, G., Thomas, A., Shahul Hameed Nabeezath Beevi, R., Gonzalez, J., Khan-delwal, K., Iacobacci, I., Benechehab, A., et al. Large language models orchestrating structured reasoning achieve Kaggle grandmaster level. *arXiv*, pp. arXiv–2411, 2024.
- Guo, S., Deng, C., Wen, Y., Chen, H., Chang, Y., and Wang, J. DS-Agent: Automated data science by empowering large language models with case-based reasoning. *arXiv preprint arXiv:2402.17453*, 2024.
- He, J., Ghosh, R., Walia, K., Chen, J., Dhadiwal, T., Hazel, A., and Inguva, C. Frontiers of large language model-based agentic systems-construction, efficacy and safety. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM*, pp. 5526–5529, 2024.
- Hong, S., Lin, Y., Liu, B., Liu, B., Wu, B., Zhang, C., Wei, C., Li, D., Chen, J., Zhang, J., et al. Data interpreter: An LLM agent for data science. *arXiv preprint arXiv:2402.18679*, 2024.
- Jing, L., Huang, Z., Wang, X., Yao, W., Yu, W., Ma, K., Zhang, H., Du, X., and Yu, D. DSBench: How Far are Data Science Agents from Becoming Data Science Experts? *arXiv preprint arXiv:2409.07703*, 2024.
- Maojun, S., Han, R., Jiang, B., Qi, H., Sun, D., Yuan, Y., and Huang, J. A survey on large language model-based agents for statistics and data science. *The American Statistician*, pp. 1–21, 2025.
- Qiu, J., Lam, K., Li, G., Acharya, A., Wong, T. Y., Darzi, A., Yuan, W., and Topol, E. J. LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420, 2024.
- Wang, X., Chen, Y., Yuan, L., Zhang, Y., Li, Y., Peng, H., and Ji, H. Executable code actions elicit better LLM agents. In *Forty-first International Conference on Machine Learning, ICML*, 2024.

⁷<https://streamlit.io>