

Template: AI Model Research Documentation Sheet (AIRDocS)

Version 1 – December 2024

Researcher(s):

Project:

Research Setup

- **Research Question:** [Your specific research question and how language models contribute to the answering of the question]
- **Methodology:** [Your methodology]

Dataset¹

- **Dataset Purpose:** [Describe the purpose of this dataset (e.g., research dataset).]
- **Source Data:** [Describe the source data (e.g., historical newspapers).]
- **Source Collection:** [Describe how the source data was collected.]
- **Licensing Information:** [Does your dataset contain licensed restricted data? Name the license and what this means for the usage of the dataset.]
- **Data Fields:** [List and describe the fields present in the dataset. Mention their data type, and whether they are used as input or output in any of the tasks the dataset currently supports. If the data has span indices, describe their attributes, such as whether they are at the character or word level, whether they are contiguous or not, etc. If the dataset contains example IDs, state whether they have an inherent meaning, such as a mapping to other datasets or pointing to relationships between data points.]
- **Data Size:** [Describe how big the dataset is.]
- **Languages:** [Describe the Language or languages of the dataset.]
- **Data Splits:** [Describe and name the splits in the dataset if there are more than one. Describe any criteria for splitting the data, if used.]
- **Dataset Version:** [Provide information on the version of the dataset.]
- **Personal and Sensitive Information:** [Does your dataset contain sensitive or personal information, such as identity data, demographic data or biometric data.]

¹ For the dataset description, following templates were used as reference: H., Claeysens, S., Colavizza, G., Freire, N., Irollo, A., Lehmann, J., Neudecker, C., Osti, G., & van Strien, D. (2023). Datasheets for Digital Cultural Heritage Datasets. Zenodo. Published September 25, 2023, Version 1. <https://zenodo.org/records/8375034>; Hugging Face Dataset Card Creation Guide, https://github.com/huggingface/datasets/blob/main/templates/README_guide.md; Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. <https://doi.org/10.1145/3458723>.

- **Discussion of Bias:** [Provide descriptions of specific biases that are likely to be reflected in the data.]

Data Processing

Input Data

- **Preprocessing:** [Steps taken for preprocessing for all fields where applicable (e.g., reducing context window for the model analysis or removing break lines or extra spacing for evaluation.)]
- **Ground Truth:** [Is ground truth included in the dataset? How is the ground truth formatted (e.g., json, xml...)?]
- **Ground Truth Size:** [How much ground truth was created?]

Output Handling

- **Output Format:** [What format has the output?]
- **Post-processing:** [Any processing of model outputs]
- **Storage:** [How outputs are stored]

Model Identification

Model Details

- **Model Name:** [e.g., GPT-4, LLaMA-2-70b]
- **Version/Checkpoint:** [Specific version used]
- **Access Method:** [API/Local deployment]
- **Provider/Source:** [e.g., OpenAI, Hugging Face]
- **Access Date:** [When you started using this version]

Model Selection model Rationale

[Brief explanation of why this was chosen for your research (e.g., advantages over alternative models, covers specific languages, technical fit with research goals, sustainability reasons, size, and specific capabilities that address your use case requirements.)]

Technical Setup

- **Implementation Method:** [How you're accessing/running the model]
- **Computing Environment:**
 - Hardware: [GPU/CPU specifications]
 - Software: [Framework versions, dependencies]
 - Runtime: [Cloud/Local/Cluster]

Model Configuration

Parameter Settings

- Temperature: [X]
- Top_p: [X]
- ...

Rationale for Settings

[Explain why these specific parameters were chosen]

Prompt/Input Design

Prompt/Input Template

[e.g., your prompts]

Prompt/Input Development

- **Iterations:** [Major iterations in prompt/input development]
- **Testing Process:** [How prompts/inputs were tested]
- **Final Selection:** [Why this prompt/input was chosen]

Validation Methodology

Quality Control

- **Validation Method:** [How outputs are validated]
- **Acceptance Criteria:** [Criteria for valid outputs]
- **Error Handling:** [How errors are handled]

Reproducibility Measures

- **Random Seeds:** [If used, what seed values were used to ensure reproducibility of the random processes? How were these seeds applied across different runs or experiments?]
- **Control Measures:** [What steps were taken to maintain consistency across all experiments? This should include model parameters, computing environment, and input formatting choices. Were these settings documented before starting the experiments.]
- **Verification Steps:** [What verification process was used to ensure result quality and reliability, including the model's self-verification capabilities (like self-consistency checks), automated validation procedures (like multiple runs), and human oversight (such as expert review)? Document who verified what, when, and how.]

Performance Analysis

Metrics

[What metrics were used to evaluate the model's performance in terms of accuracy (including both success and error rates), efficiency (including processing speed and resource usage), and practical considerations (such as computational or API costs)? For each metric, specify how it was measured, what thresholds were considered acceptable, and how these measurements inform the research conclusions.]

Limitations

- **Observed Issues:** [Document specific technical and measurable performance deficiencies, including: quantitative metrics below target thresholds, throughput bottlenecks, resource utilization problems, response time variations, accuracy rates for

different tasks, error rates in specific scenarios, and reproducible failure cases. Focus on concrete, measurable aspects rather than qualitative observations.]

- **Workarounds:** [Describe technical solutions implemented to improve model performance metrics, such as: model parameter adjustments, optimization techniques, caching strategies, parallel processing implementations, hardware configuration changes, and system architecture modifications. Include specific configuration changes, code adjustments, and infrastructure modifications that led to measurable improvements.]
- **Unresolved Challenges:** [Detail remaining technical limitations with quantifiable impact, such as: maximum input size constraints, minimum latency bounds, memory usage limits, accuracy ceilings for specific tasks, and resource requirements that couldn't be reduced further. Focus on technical barriers that can be measured and benchmarked, rather than qualitative insights or research directions.]

Resource Usage

Computational Resources

- **Processing Time:** [Time taken for processing]
- **Memory Usage:** [Peak memory requirements]
- **Storage Requirements:** [Storage needed]

Cost Analysis

- **Total Cost:** [If applicable]
- **Cost Breakdown:** [By component/stage]
- **Optimization Efforts:** [Cost-saving measures]

Ethical Considerations

Bias Assessment

- **Identified Biases:** [List and analyze any observed biases in model outputs, such as language bias, representation of different demographics, domain-specific prejudices, or content moderation inconsistencies. Include specific examples where possible.]
- **Mitigation Steps:** [Detail the strategies implemented to address identified biases, including prompt engineering techniques, content filtering methods, output validation procedures, and any additional safeguards put in place to ensure fair and balanced model responses.]
- **Remaining Concerns:** [Describe any persistent biases or ethical challenges that couldn't be fully resolved through mitigation efforts. This includes potential impacts on different user groups, limitations in certain use cases, and recommendations for responsible usage given these known issues. Consider both immediate and potential long-term effects of deploying the LLM in your specific context.]

Privacy and Security

- **Data Privacy:** [Detail the measures implemented to protect personal and sensitive information when using LLMs [e.g., using models locally].]
- **Compliance:** [Outline how the LLM implementation adheres to relevant data protection regulations (such as GDPR), industry standards, and organizational

policies. Include documentation of compliance assessments, auditing procedures, and any specific requirements for your use case or jurisdiction.]

Research Notes

- [Document significant behavioral characteristics of the model observed during research, including response patterns, performance variations across different tasks, limitations in specific contexts, and any consistent strengths or weaknesses identified through systematic testing.]
- [Detail surprising or counterintuitive findings that emerged during the research process, such as unexpected model capabilities, limitations that weren't anticipated, or interesting interactions between different experimental variables that weren't part of the original research questions.]
- [Summarize key learnings that could inform future research directions, including methodological improvements, promising areas for further investigation, potential pitfalls to avoid, and specific hypotheses or questions that emerged from this work but weren't fully explored. Include both technical insights about model behavior and practical recommendations for research design.]

Documentation

Code Repository

- **Location:** [Link to code repository]
- **Version:** [Code version/commit]
- **Dependencies:** [Required packages/versions]

Data Storage

- **Location:** [Where data is stored]
- **Format:** [File formats used]
- **Access:** [How to access the data]

Supporting Materials

- [Links to relevant documentation]
- [Additional resources]
- [Reference materials]

Bibliography

Gebu, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92. <https://doi.org/10.1145/3458723>

H., Claeysens, S., Colavizza, G., Freire, N., Irollo, A., Lehmann, J., Neudecker, C., Osti, G., & van Strien, D. (2023). Datasheets for Digital Cultural Heritage Datasets. Zenodo. Published September 25, 2023, Version 1. <https://zenodo.org/records/8375034>; Hugging Face Dataset Card Creation Guide, https://github.com/huggingface/datasets/blob/main/templates/README_guide.md.

Sarah Oberbichler, Leibniz Institute of European History, ORCID: <https://orcid.org/0000-0002-1031-2759>

Liang, P., Bommasani, R., Lee, T., et al. (2023). Holistic Evaluation of Language Models. Transactions on Machine Learning Research. <https://openreview.net/forum?id=iO4LZibEqW>

Smith, G. R., Bello, C., Bialic-Murphy, L., et al. (2024). Ten simple rules for using large language models in science, version 1.0. PLOS Computational Biology, 20(1), e1011767. <https://doi.org/10.1371/journal.pcbi.1011767>.

Wagner, S., Muñoz Barón, M., Falessi, D., & Baltes, S. (2024). Towards Evaluation Guidelines for Empirical Studies involving LLMs. arXiv:2411.07668v2 [cs.SE]. <https://arxiv.org/abs/2411.07668>.

Watkins, R. Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. *AI Ethics* 4, 969–974 (2024). <https://doi.org/10.1007/s43681-023-00294-5>.