

GUIDE-LLM: A checklist for reporting studies with large language models in behavioral and social science

The GUIDE-LLM checklist provides a standardized framework for reporting studies that use large language models (LLMs) in the behavioral and social sciences. It aims to promote transparency, reproducibility, and ethical accountability across all stages of LLM-based research.

How to complete the checklist:

Fill in each item with brief, specific information about how LLMs were used in your study. Where an item does not apply, write “N/A” and, if helpful, note why. If multiple LLMs were used for different purposes, complete the relevant sections separately for each model. You may refer to sections or appendices in your manuscript rather than repeating text.

Scope of LLM use	Answer
<p>Item A.1: LLMs were used in this project for:</p> <p>Explanation: Briefly describe how and for what purposes LLMs were used in the study. This may include one or several stages of the research workflow, depending on the project’s design and aims. The following examples illustrate common categories:</p> <ul style="list-style-type: none">● Research design (e.g., hypothesis generation, literature search, or creating surveys/stimuli).● Data processing (e.g., transcription, translation, or data extraction).● Analysis (e.g., data labeling, summarization, pattern detection, or code writing).● LLM as research object (e.g., studying LLM behavior, benchmarking LLMs, or bias assessment of LLMs).● Participant-facing settings (e.g., LLM used as an intervention, studying human interactions with LLM chatbots).● Communication (e.g., paper writing, editing, or reviewing). <p>Depending on the specific use case described here, different checklist items may later be relevant, and, in many cases, it may be necessary that later items in the checklist are reported separately for each use case.</p>	<p>An LLM pipeline was used for multi-label text classification of 3.7 million official development assistance (ODA) project descriptions into 17 disease categories, adapted from Level 2 causes in the Global Burden of Disease (GBD) hierarchy. The LLM classifications were then used to (i) aggregate disease-specific ODA by country and year, and (ii) compare disease-specific ODA with disease burden (DALYs) at the country level to quantify aid–burden misalignment.</p>
<p>Item A.2: Degree of automation (human-in-the-loop vs. fully automated):</p> <p>Explanation: Indicate how much human oversight was involved. For example, was each LLM output reviewed or edited by a person, or was it used automatically? For participant-facing tasks, state whether humans checked outputs before showing them to participants or whether participants interacted with the LLM directly. Specify who provided oversight (e.g., student assistant, expert, PI).</p>	<p>The disease classification pipeline was largely automated (except for the human validation).</p>

Model/system details	Answer
<p>Item B.1: Model name, including provider, model size, exact version/ID, date of access, and source link (if possible):</p> <p>Explanation: Report the exact model names (including provider, version, and date accessed). Avoid generic labels like “ChatGPT” or “GPT-4”; instead, use detailed model names such as “GPT-4o-mini-2024-12-17 (OpenAI)” or “Llama-3.1-8B (Meta; accessed via HuggingFace in May 2025)”. For locally deployable models, please also enter a source link (e.g., the URL to the HuggingFace page). If you tested multiple models, it is encouraged to name them and briefly explain which one you used in the final study and why.</p>	<ul style="list-style-type: none"> ● Model name: Meta-Llama-3.1-70B-Instruct-Turbo ● Provider: Meta (accessed via Together AI) ● Model size: 70B parameters ● Exact version: Meta-Llama-3.1-70B-Instruct-Turbo ● Date of access: June 2025 ● Source link: https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
<p>Item B.2: Model access (e.g., API, web interface, local) and context mode (e.g., chat mode or separate calls):</p> <p>Explanation: Note how you accessed the models (e.g., API, web interface, local installation) and whether you used LLMs in chat mode (ongoing conversation) or stateless mode (separate prompts). Mention the exact API name and version, since different access modes may influence responses (e.g., due to differences in model routing).</p>	<ul style="list-style-type: none"> ● Access mode: API access via Together AI for Meta-Llama-3.1-70B-Instruct-Turbo ● Context mode: Each project description was passed as an independent classification request together with a fixed prompt describing the task and label set (stateless mode).
<p>Item B.3: Relevant LLM configurations reported (as applicable), such as temperature, max tokens, seed, and number of runs:</p> <p>Explanation: List any configuration settings that affect outputs, such as:</p> <ul style="list-style-type: none"> ● temperature (which controls model randomness) ● top_k, top_p, max tokens (which controls sampling so that, e.g., only the <i>k</i> most probable tokens are considered, or to enforce a length limit) ● Certain penalties that discourage repetition (e.g., a frequency penalty to reduce the likelihood of tokens proportional to how often they have already appeared; a presence penalty reduces the likelihood of any token that has appeared at least once) ● Stop sequences (which halt generation when such a top sequence is produced, such as, e.g., ["\n\n", "END"]) ● Number of completions or runs (which is often used to capture variance in outputs across repeated generations) ● Quantization level (e.g., FP16, INT8, INT4) to change numerical precision beyond the default ● Reasoning-related settings, such as whether a specific structured reasoning is enabled, the specified reasoning effort level (e.g., low/medium/high or numerical settings that influence the depth of the reasoning), and any compute or inference budget constraints tied to the chosen reasoning mode 	<ul style="list-style-type: none"> ● Temperature: 0.7 ● Top_k: 50 ● Top_p: 0.7 ● Max tokens: 40 ● Repetition penalty: 1.0 ● Quantization level: FP8 ● Reasoning: n/a

<p>Item B.4: Customization:</p> <p>Explanation: Check and describe any modifications or extended capabilities your setup used. Examples: fine-tuning (e.g., via LoRA; Low-Rank Adaptation) is used to adapt a pretrained model to domain materials; retrieval-augmentation generation (RAG) is a technique in which the model retrieves relevant information from external sources (e.g., databases). Here, web search refers to whether the LLM could retrieve other information from the web; automated prompt optimization refers to certain wrappers (e.g., DSPy) that treat prompts as a trainable parameter; agentic workflows refer to multi-step reasoning or delegated actions that go beyond simple tool/function calling. (e.g., via LangChain, AutoGPT, CrewAI). For post-training, describe any custom refinement processes applied to the LLMs, including alignment methods or model-level optimization techniques used to adjust behavior after pretraining (e.g., reinforcement learning from human feedback (RLHF), direct preference optimization (DPO)). Here, the goal is to specify any added customizations and any provider-specific characteristics that meaningfully shape system behavior in order to enable others to understand and accurately reproduce your setup.</p>	<ul style="list-style-type: none"> <input checked="" type="checkbox"/> Base model <input type="checkbox"/> Fine-tuning <input type="checkbox"/> RAG (retrieval-augmented) <input type="checkbox"/> Automated prompt optimization <input type="checkbox"/> Tool/function calling <input type="checkbox"/> Web search <input type="checkbox"/> Agentic workflows <input type="checkbox"/> Other adaptations (e.g., safety mechanisms) <p>Description: n/a</p>
<p>Item B.5: Did the LLM session(s) include persistent memory across interactions?</p> <p>Explanation: Indicate whether the LLM could “remember” previous conversations (i.e., had persistent memory). Unless such memory is disabled, there may also be spillover effects from other chat windows or prior conversations, which can influence outputs even when not intended.</p>	<ul style="list-style-type: none"> <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> N/A

Prompts	Answer
<p>Item C.1: Exact prompt(s) reported:</p> <p>Explanation: Whenever possible, include the exact text of prompts you used, including in-context examples or demonstrations provided to the LLM. Even small wording changes, formatting, or ordering of examples can substantially affect outputs. If full prompts cannot be shared (e.g., due to privacy or length), include a redacted or representative example or link to the full prompt in a repository (e.g., OSF, GitHub).</p>	<p>See Supplementary Materials S3</p>
<p>Item C.2: System-wide instructions (if any):</p> <p>Explanation: Note any <i>system-level</i> instructions that guide the model’s general behavior (e.g., “You are a helpful assistant.”). These are commonly not directly visible but can be accessed through the API.</p>	<p>n/a</p>

Data inputs & privacy	Answer
<p>Item D.1: Handling of personal or sensitive data (if any) (e.g., consent for data processing):</p> <p>Explanation: If any personal, sensitive, or identifiable data were processed, describe how they were handled in compliance with ethical standards and data protection laws. Researchers should indicate whether participants explicitly consented to their data being analyzed with an LLM, particularly when proprietary, cloud-based models are used. Such processing typically involves transferring data to a private company that may retain them indefinitely, which raises additional ethical and legal considerations. Beyond consent, describe how sensitive or identifiable data were handled (e.g., de-identification, anonymization, masking) and whether the LLM provider offers safeguards such as excluding inputs from training or storage. Clarify where data were stored or processed and how applicable legal/ethical requirements were met. If relevant, address cross-border transfers, as data may be stored in jurisdictions with different privacy laws (e.g., EU vs. US), with implications for compliance with GDPR, HIPAA, or other frameworks.</p> <p>For context, some providers (e.g., OpenAI) may log or inspect prompts even when the data are not used for model training. For sensitive datasets, zero-retention configurations may be required (e.g., the MIMIC datasets can only be used with OpenAI models if a zero-retention checkpoint is enabled).</p>	<p>Our work did not involve any personal or sensitive data.</p>

Validation & interpretation	Answer
<p>Item E.1: Human validation of LLM outputs:</p> <p>Explanation: Describe whether and how human reviewers examined the model's outputs, and the degree of independence they had in doing so. Specify the reviewers' roles (e.g., domain experts, research assistants, subject-matter specialists) and relevant expertise, as well as how many reviewers participated and how their work was organized. Indicate whether outputs were independently annotated, double-checked by multiple reviewers, or merely approved or edited post-hoc by a lead author or investigator.</p> <p>Clarify what dimensions of performance were examined. These may include known performance metrics from ML/AI such as accuracy or other metrics like citation correctness, hallucination detection, agreement or inter-rater reliability. State whether qualitative judgments, quantitative metrics, or both were used. If outcome assessment required subjective interpretation, describe assessor qualifications, instructions provided, relevant demographics, and any inter-assessor agreement measures.</p> <p>Describe the selection procedure for the reviewed outputs—whether all outputs were examined, a random sample was drawn, or specific cases (e.g., rare events or high-stakes responses) were oversampled to capture potential rare or critical errors. Further report how reviewers were trained or instructed, what criteria or rating scales they used, and how disagreements were resolved. For multi-reviewer settings, provide any inter-rater or inter-assessor reliability statistics (e.g., Cohen's κ or Krippendorff's α). Finally, note whether reviewer feedback was used purely for validation or also to refine prompts, retrain models, or adjust study procedures.</p>	<p><input checked="" type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> N/A</p> <p>Description: LLM outputs were validated against a manually annotated dataset of 2,060 ODA projects labeled independently by two members of the research team. Agreement was evaluated using overall accuracy (0.93), micro-averaged F1-score (0.94), macro-averaged F1-score (0.93), and per-disease precision and recall (> 0.9 for most disease categories), alongside robustness analyses using Cohen's kappa (κ) and label consistency rate (LCR) across subsets and model variants. The 2,060 projects were sampled from the full dataset, with additional targeted subsets to test failure modes: long descriptions, descriptions with negation, and short, ambiguous, or non-English descriptions. Details are in Supplementary Materials S1.2.</p>
<p>Item E.2: Describe any relevant post-processing (e.g., filtering in case of format mismatches, unit conversions, etc.):</p> <p>Explanation: Describe any steps you took to clean or reformat LLM outputs (e.g., converting "positive/neutral/negative" to numeric codes, handling missing values, removing malformed entries). State how you handled inconsistent or unusable outputs and whether corrections were made with an automated script or manually. For example, when generating quantitative estimates (e.g., word counts, probabilities, or durations), the model may return values embedded in free text (e.g., "3.5 seconds") that require parsing and conversion into standardized numerical units. Post-processing steps should be described clearly, including how formatting errors, null responses, or inconsistent output structures were handled, whether automated scripts or manual corrections were used, and whether any data were excluded or reinterpreted as a result.</p>	<p>LLM outputs (i.e., lists of possibly multiple disease labels from the 17 disease categories plus fallback labels "Other" and "General Health") were parsed programmatically. For multi-label projects, funding amounts were split equally across all assigned diseases to prevent double counting while reflecting cross-cutting projects. Multi-year projects relied on CRS year-specific disbursements, which were used as reported and assigned to the corresponding years. Post-processed labels and allocations were aggregated to obtain disease- and country-specific ODA per year and per capita (using World Bank population data), then matched to DALYs for correlation and misalignment analysis.</p>

Reproducibility	Answer
<p>Item F.1: Code/notebooks/scripts for LLM calls shared:</p> <p>Explanation: Indicate whether you have shared materials such as code, prompts, logs, or transcripts. Make sure sensitive information (e.g., API keys, private data) is removed. For code, make sure to add a README file.</p>	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> N/A Link/DOI: https://github.com/forsterkerstin/tracking-health-aid-disparities

Competing interests	Answer
<p>Item G.1: Funding, support, or other relevant relationships (including in-kind access to compute or models, or professional affiliations):</p> <p>Explanation: Disclose any current or past funding, support, or other relevant relationships with entities that have a financial interest in LLMs (this includes not just AI companies like OpenAI, Anthropic, but also tech companies developing or investing in AI, e.g., Google, Meta, Microsoft). This could include (but is not limited to): research funding from or collaborative research with a company with an interest in LLMs for this project or any other project within the past years; in-kind access to compute or models; current or former professional affiliations with a company with an interest in LLMs; personal investments (e.g., stocks) in companies with an interest in LLMs; familial relationship with an employee of a company with an interest in LLMs; etc. Disclose these relationships regardless of whether or not you believe they impacted the research.</p>	<input type="checkbox"/> Yes. Description: <input checked="" type="checkbox"/> No

Optional items	Answer
<p>Justification for the LLM choice along the following dimensions:</p> <p>Explanation: Briefly explain why you chose this model for your study and thereby explain why the selected model is suited to the research needs. Consider:</p> <ul style="list-style-type: none"> • Performance: Explain why this model's accuracy or size was appropriate for the research task, and, if relevant, consider compute cost, scalability, and real-time feasibility. • Transparency: Was the model open-weight or closed-weight? Do we know what training data the model was exposed to or what the training objective was? If not, what are the implications for interpreting results? In terms of openness, different levels can be distinguished, in which only the model weights are publicly available ("open-weight") versus LLMs where the entire training process is available to allow for further customization and transparency of the LLM itself ("open-source"). • Reproducibility: Can the model be shared (so results can be reproduced in the future), or is access limited to proprietary/cloud services (so the access to the model may be altered, restricted, or discontinued)? • Ethical considerations: Are there ethical reasons (including safety) for the model choice, such as running a model locally to avoid sharing sensitive data? • Other determinants can be the cost or ease-of-use (e.g., due to accessibility of certain models via university infrastructures), if their benefits outweigh trade-offs in other dimensions. 	<input checked="" type="checkbox"/> Performance – Description: We selected Meta-Llama-3.1-70B-Instruct because it was a frontier open-weight LLM at the time of writing and demonstrated strong performance on structured text classification tasks. <input type="checkbox"/> Transparency – Description: <input checked="" type="checkbox"/> Reproducibility – Description: The model is open-weight, and, therefore, our pipeline is fully reproducible. <input type="checkbox"/> Ethical considerations – Description: <input type="checkbox"/> Others (e.g., cost, ease-of-use): – Description:

<p>Discussion of the rationale for the prompt design:</p> <p>Explanation: Explain how you designed your prompts. For example, did you use a structured format, follow existing design guidelines, or use an automated prompt-optimization tool?</p>	<p><input checked="" type="checkbox"/> Yes. Description: The prompt design followed best practices in prompt engineering and prior research (Feuerriegel et al., 2025, Giray et al., 2023, Lin et al., 2024), and includes a task description, a list of possible labels, and the description of the ODA project to be classified. The model was instructed to adopt a conservative classification strategy to reduce false positives, using fallback categories (“Other” and “General Health”) where applicable.</p> <p><input type="checkbox"/> No <input type="checkbox"/> N/A</p>
<p>Comparison against other methods/LLMs:</p> <p>Explanation: If you compared different models, prompts, or methods, describe how you made the comparison and what criteria you used (e.g., accuracy). State whether you used the same inputs across methods and how you ensured fair comparison. If you considered but excluded alternative methods/LLMs, briefly explain why. If such comparisons are made, report the exact benchmarking procedure and the level of agreement between methods. If other LLMs or methods were considered but not used, briefly explain the rationale (e.g., cost, access, performance, or ethical considerations).</p>	<p><input checked="" type="checkbox"/> Yes. Description: We compared against an alternative LLM (gpt-4.1-mini-2025-04-14) and report, e.g., Cohen’s kappa. Details are in Supplementary Materials S2.</p> <p><input type="checkbox"/> No <input type="checkbox"/> N/A</p>
<p>Training data leakage risks addressed:</p> <p>Explanation: Note whether you considered the chance that your evaluation materials (e.g., test data, questionnaires) were already part of a model’s training data. If relevant, describe steps taken to mitigate or detect leakage, such as novelty checks, re-phrasing or re-structuring items, or using hold-out materials not widely available.</p>	<p><input type="checkbox"/> Yes <input type="checkbox"/> No <input checked="" type="checkbox"/> N/A</p>

Potential risk of bias or other systematic differences in LLM behavior that could affect the study's conclusions?

Explanation: Reflect on whether the LLM might perform differently across groups, languages, or contexts in ways that could affect your conclusions. If you checked for or mitigated such bias, describe how (e.g., subgroup analyses, balanced sampling, sensitivity checks). Where applicable, researchers may consult dedicated bias and fairness checklists for structured guidance on identifying and addressing these issues.

Yes. Description: There is a potential risk that the LLM may exhibit systematic differences in performance across languages, text structures, or regional reporting practices, which could in turn influence disease classifications and thus affect downstream measures of aid–burden alignment. Project descriptions in the CRS vary widely in length, clarity, and linguistic quality, and many were originally written in non-English languages. Although all non-English descriptions were translated prior to classification, machine translation may introduce inconsistencies that could differentially affect projects from specific donor countries or regions. Likewise, shorter or less detailed descriptions may lead to lower classification accuracy for certain disease areas. To mitigate these risks, we implemented several safeguards through extensive validation. We offer a detailed explanation of potential limitations in the Section “Discussion”.

No

Conversation transcripts:

Explanation: For studies where participants or researchers interacted directly with an LLM (e.g., chat-based interventions, behavioral experiments, or coding tasks), provide anonymized transcripts or representative examples of the interactions. Redact any personally identifiable or sensitive information.

Yes, shared without sensitive information. Location:

No. Reason for not sharing:

N/A

Discuss relevant ethical implications of the research:

Explanation: Briefly discuss the broader ethical considerations of your study beyond procedural compliance (e.g., IRB approval). Address potential risks and safeguards, including:

- **Participant well-being:** foreseeable harms (emotional, reputational, financial, environmental), inclusion/exclusion criteria, effects on vulnerable groups, accessibility, use of deception and debriefing, autonomy and informed consent (including clear disclosure of AI involvement).
- **Fairness and equity:** whether outputs or outcomes may reflect or amplify bias or discrimination; note any subgroup analyses or mitigation strategies.
- **Safety:** broader risks (such as from agentic or autonomous behavior), or safeguards for harmful or adverse outputs (e.g., moderation layers, content filters, human-in-the-loop review, escalation protocols, red-teaming, dual use).

Since our study relies exclusively on publicly available data on development aid projects, the risks to individuals or identifiable groups are minimal. No personal, sensitive, or confidential information is processed. Nonetheless, several broader ethical considerations are relevant. First, our LLM pipeline should be used for monitoring, and we caution for any use related to directly optimizing aid allocation. We thus refrain from interpreting our results as prescriptive evidence. Second, fairness and equity considerations arise because systematic differences in LLM behavior—such as variation in performance across languages, text lengths, or writing styles—could inadvertently bias disease classifications. We address these risks through translation, preprocessing, manual validation, and comparison against keyword-based baselines, and we interpret aid–burden misalignment strictly as a descriptive metric rather than a prescriptive judgment about funding adequacy. Finally, the overall societal impact of the study is expected to be positive. By improving transparency in global health financing and enabling scalable, reproducible monitoring of health-related ODA, our approach can support more equitable and evidence-based allocation of resources.

Computational resources (e.g., API call counts, tokens, financial costs, or compute time):

Explanation: Report how much computing power or cost your study required (e.g., the number of GPUs, their model, and the runtime), ideally at both study-wide and, if applicable, per-participant levels.

EUR ~500 via Together AI. All other LLM operations run on a single NVIDIA A100 40GB in a reasonable timeframe (<1 week).

How to cite: Feuerriegel et al. (2026). A consensus-based reporting checklist for large language models in behavioral and social science.