A Final Codebook

Table 2. The final codebook with four code categories as well as the Inter-Rater Reliability. Multiple codes can apply to one paper.

Category	Codes	IRR
Contribution Types (7 codes)	Empirical, Artifact, Methodological, Theoretical, Dataset, Metareview, Opinion	**0.866**
Application Domain (10 codes)	Communication & Writing, Augmenting Capabilities, Education, Responsible Computing, Programming, Reliability & Validity of LLMs, Well-being & Health, Design, Accessibility & Aging, Creativity	**0.849**
Roles of LLMs (5 codes)	LLMs as system engines; LLMs as research tools; LLMs as participants & users; LLMs as objects of study; Users' perceptions of LLMs	**0.773**
Limitations & Risks (29 codes)	**Limitations on LLM Performance** LLM bias toward different groups, limited data coverage in the training data, non-deterministic response, hallucination, unspecificied errors and biases; **Limitations on Research Validity** internal and/or external validity across users, contexts, models, prompts; **Limitations on Resource** computational cost, financial cost, lack of evaluation standards; **Risks to Society, consequences** economic harms, representational harms, misinformation harms, malicious use, hate speech, and environmental harms. **Note**: *There are 22 low-level codes described in the full paper. However, we had an initial set of 29 low-level code. We had four code "others" under each coarse categories (e.g., other LLM performance issues). We also merged three other codes during the coding process (i.e., latency, lack of access to open/close models, prompt-induced performance issues). These codes overlap with the definition of the 22 main codes.*	**0.887** (**0.633** for the 29 low-level codes)

B Definition of Contribution Types

- **Empirical Contribution**: They provide findings based on observation and data-gathering, including experiments, user tests, field observations. Interviews, surveys, focus groups.
- **Artifact Contribution**: They provide news systems, architectures, tools, toolkits, techniques.
- **Methodological Contribution**: They inform us how we carry out our work.
  - Note that this contribution type focuses on research methods contribution.
  - If the paper creates a novel application or framework for users to interact with LLMs (e.g., through prompting or finetuning), it is a methodological contribution.
  - If the paper uses LLMs to brainstorm in a unique application domains, it is not methodological contribution because they largely rely on the LLMs performance.
- **Theoretical Contribution**: They consist of new or improved concepts, definitions, models, principles, or frameworks.

  - Note that theoretical contributions have to validate or develop theory.
  - Drawing from [137], if a paper applies or adopts a theory to HCI research, this is a theoretical contribution.
  - In other cases the paper's main goal should be developing a theory (e.g., design theory, ethical theory, guideline, framework, design space).
  - If a paper comes up with design requirement as a formative study or in the discussion section or design implications section, they are not theoretical contribution.
- **Dataset Contribution**: They provide a new and useful corpus, often accompanied by an analysis of its characteristics, for the benefit of the research community. Note that we only count dataset contribution if the paper open sources the data, or explicitly mention the contribution type.
- **Survey/Metareview Contribution**: Survey research contributions and other meta-analyses review and synthesize work done on a research topic with the goal of exposing trends and gaps
- **Opinion Contribution**: Opinion research contributions, also called essays or arguments, seek to change the minds of readers through persuasion. They are position papers. Note that every paper might convey some opinions, but the major contribution should be considered.