



Beyond Manual Coding: Three-Phase Validation of LLMs for Open-End Survey Responses and Exploration of Functional Integration

Allie Pierce, PhD – Director, KS&R

Keaton Wilson, PhD – Solutions Developer, KS&R

Ben Cortese, PhD – VP, KS&R

Why Researchers Need a Better Way to Evaluate AI for Open-Ends

- **Depth & Nuance:** Open-end (OE) responses contain some of the richest insight in a study, allowing researchers to uncover unexpected insights and understand the “*why*” behind responses
- **Increasing amount of OE data:** The rise of AI-based survey tools, like dynamic conversational probes, make it easier than ever to collect unstructured OE response data
- **Slow & costly to analyze:** OE coding is a highly subjective task traditionally handled by humans, making the process slow, costly, and difficult to scale.
- Researchers need approaches that **improve efficiency without compromising quality** of the results

AI provides an opportunity for more efficient OE coding workflows, but **validation is critical to ensure reliability and accuracy** of outputs



A Structured, Stage-Based Framework for Evaluation

- OE response coding workflow is typically a multi-stage process
 - Important to assess an AI-assisted OE workflow accordingly → Split evaluation into distinct stages
 - Different stages of the workflow involve different tasks, risks, and evaluation criteria
- AI does not have to succeed or fail at the entire workflow to be useful
- AI-assisted automation of subjective tasks need a hybrid approach that involves a human-in-the-loop

Codeframe Generation

- Can AI create clear, useable structure for categorizing responses?

Code Assignment

- Can AI reliably and accurately categorize OE responses into the codeframe?

Optimization and Implementation

- How much training data, context, and human involvement is optimal to reliably scale the workflow?

Stage 1: Evaluating AI for Codeframe Generation

- **Define how to assess:** What does an ‘accurate’ AI-generated codeframe look like?
 - Codeframes should be clearly defined with distinct categories and be useable for meeting the research objectives
 - AI-generated codeframes should at least match the quality of codeframes generated by human researchers.
- **Our Approach:** When assessed by researchers with subject matter expertise, how does it compare to human-generated codeframes?

Task: Generate Codeframe

3 Human OE Coding Experts & 2 AI Platforms
Same instructions & OE response data



Task: Assess Codeframes

3 separate human evaluators across 2 rounds
Blindly ranked and reviewed for accuracy & quality

AI codeframes were ranked as top
choice in **5/6 comparisons**

Stage 2: Evaluating AI for Code Assignment

- Defining “accurate” code assignment for assessing AI performance is difficult
 - Assigning OE responses into codeframe is an inherently subjective task, traditionally based on expert human judgment
- Especially true for ambiguous or nuanced OE responses:
 - A. *“The software is incredibly easy to navigate and understand – the workflow is clear”*
 - B. *“The software’s integration with other platforms is incredibly dependable, but when they go down, customer support has been unreliable at best”*
- For open-end coding, there is rarely a single perfect ground truth.
- Our Approach: **Compare AI performance against our expert human coders code assignment taking inter-rater reliability into account**
- Results: **Find out in Berlin!**

Code ID	Code Name
1	ease of use
2	feature integration
3	reliability
4	insight quality
5	support feedback
6	nonsense or off-topic
7	none/na/missing

Ongoing Evaluation & Tradeoffs

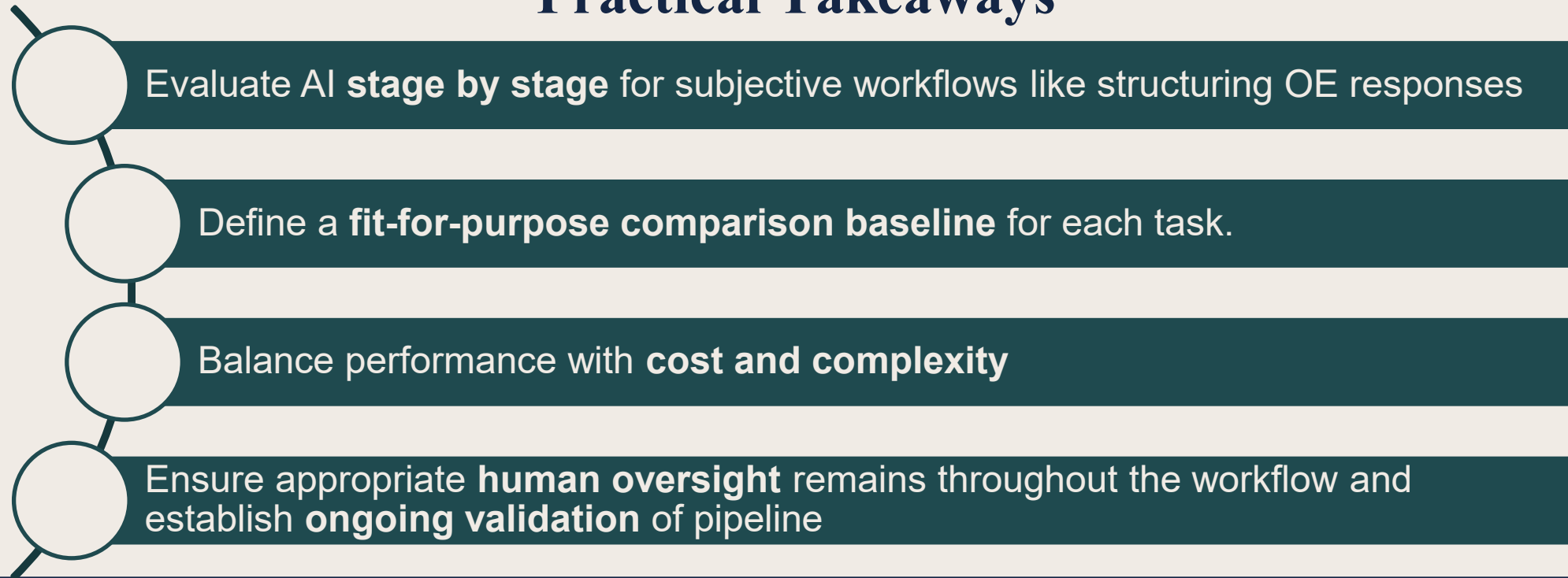
- AI performance comparable to traditional coding is **not** the end of the evaluation
- AI-assisted workflows require **ongoing validation**
 - Initial evaluation establishes viability of AI-assisted workflow
 - **Human in the loop** is critical
 - Need regular cadence for review and adjustment still needed due to drift
- Assess **tradeoffs of complexity, training data, and cost** to optimize workflow
 - When does added complexity stop paying off?
 - How much context or training data is worth adding?
 - What are the cost implications?

Our research shows that more context and examples can improve performance, but **more is not always better** (more on this in Berlin!)

What This Means in Practice

A **structured evaluation approach** can help researchers decide **when and how** AI is can be leveraged to assist or replace traditional OE coding methods.

Practical Takeaways



KS&R

Thank You!