03 June 2026

From Pilot to Practice:

A Framework for Testing AI Tools in Evaluation

www.glocalevalweek.org

Introductions

Dr. Kari Nelson

Technical Director, Social Impact

Julian Glucroft

Director, Dept. of Policy and Evaluation,
Millennium Challenge Corporation

Mateusz Pucilowski

Chief Executive Officer, Social Impact

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

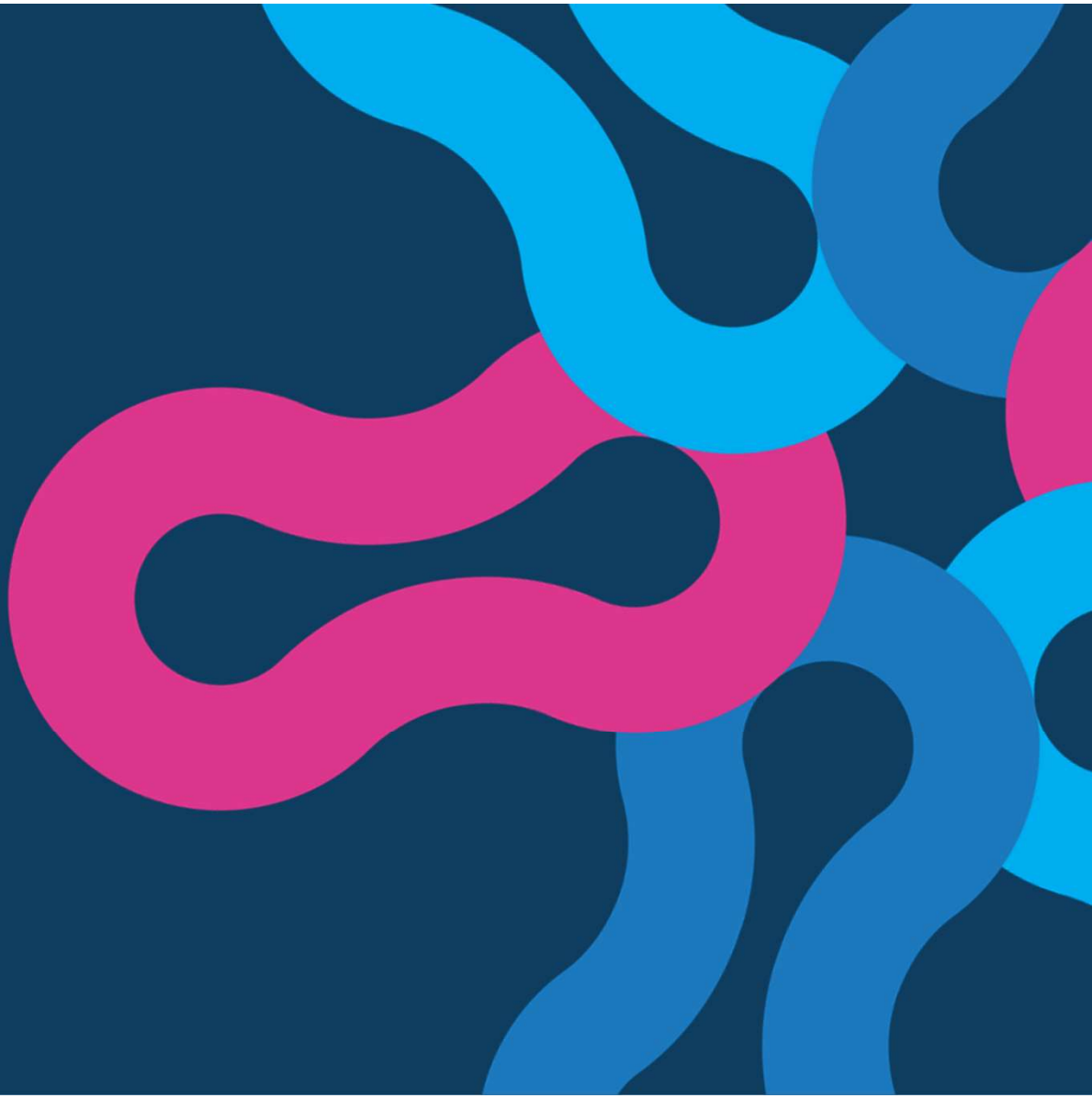
Agenda

- Context
- De-Identification
- Transcription & Translation
- Qualitative Coding
- Closing Remarks

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Context



How do we decide when AI is good enough to use in evaluation?

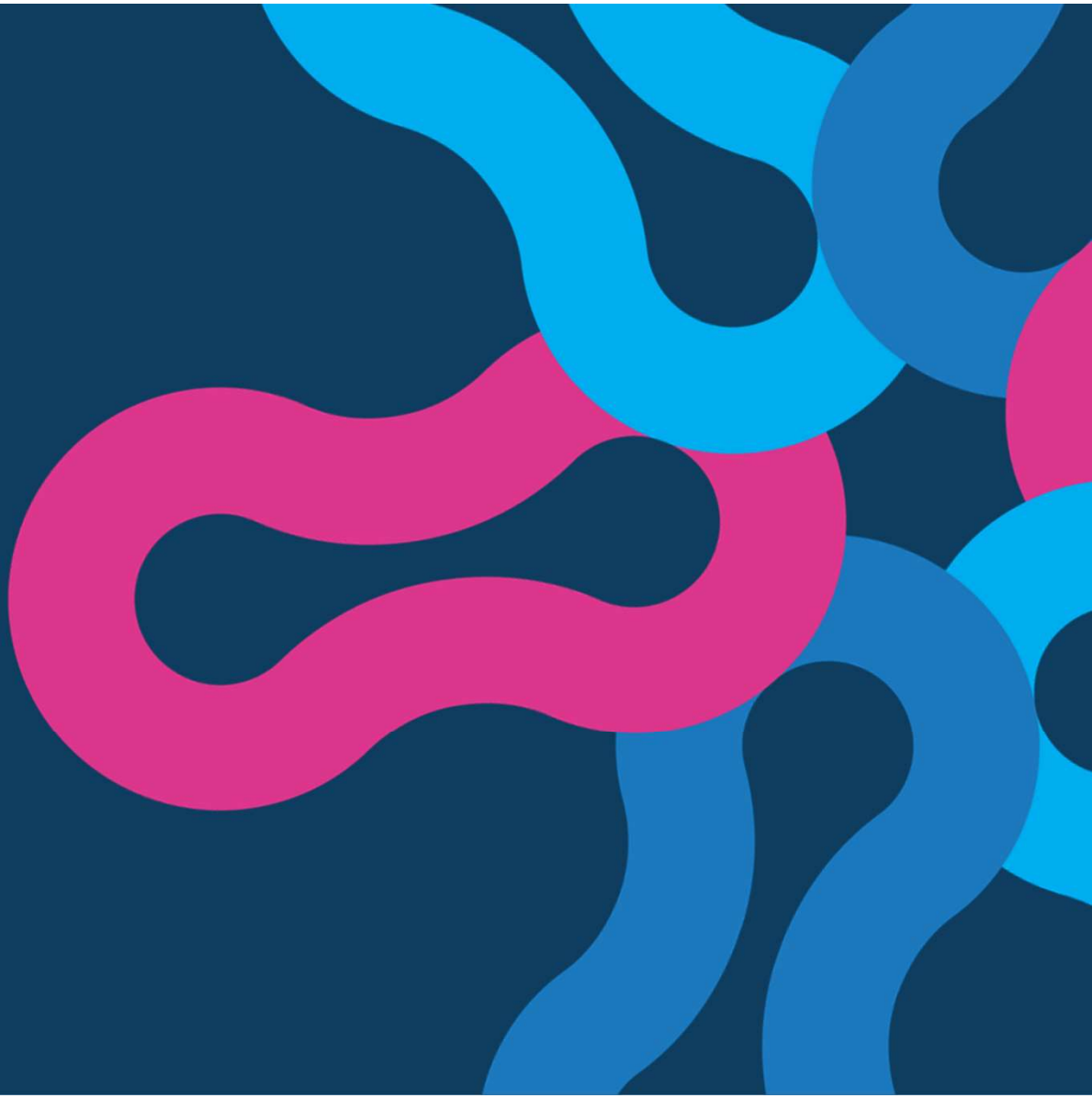
SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Piloting Overview

- Qualitative Data De-Identification
- Transcription & Translation
- Qualitative Coding

Qualitative Data De-Identification



Data De-Identification

Pilot Process

- Used ChatGPT (chat, project, custom GPT)
- Methodically tested numerous prompts and provided examples to help guide the LLM

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Data De-Identification

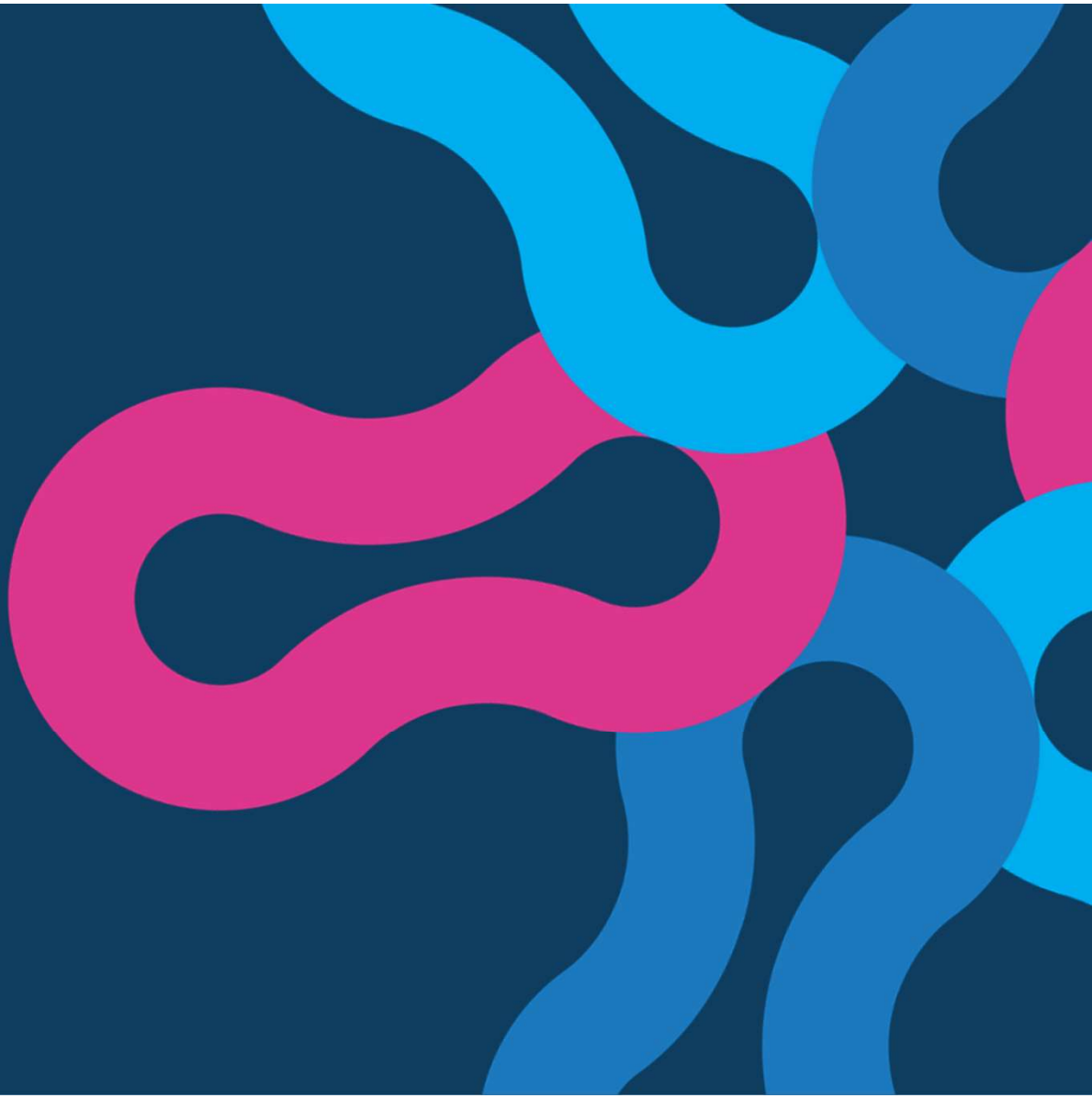
Pilot Results:

- Performance was inconsistent:
  - Sometimes identifying entire paragraphs rather than specific data points
- Decided to revert to manual deidentification
  - Transcripts already had most direct identifiers removed

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Transcription & Translation



Transcription

Pilot Process:

- Tested 3 tools: MaxQDA, Sonix, and Happy Scribe
- Using a standardized scoring matrix

Quality Criteria

-Completeness

-Additions

-Accuracy

-Management of Overlapping Speech

Platform Criteria

-Cost

-# of Languages Covered

-Security Protocols

-Transcript Editing Options

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Transcription

Pilot Results:

happyscribe 8.3/10 Performed the best (also cheapest)

SONIX 8.3/10 Performed well (but more inferences)

MAXQDA 7.6/10 Ok quality (no translation option)

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Translation

Pilot Process:

- Tested 2 tools: Sonix and Happy Scribe
- Used a standardized scoring matrix

Quality Criteria

-Completeness

-Additions

-Accuracy

-Flow

Platform Criteria

-Cost

-# of Languages Covered

-Security Protocols

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Translation

Pilot Results:

happyscribe 9.5/10 Performed the best (also cheapest)

sonix 9.4/10 Performed well (but more inferences)

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Translation

Translation Pilot Results:

happyscribe 9.5/10 Performed the best; also cheapest

sonix 9.4/10 but more inferences; struggled with accents

AI tools offer significant efficiencies, even w/o 100% accuracy

- Saved $30,000+ on MCC Niger Evaluation

Piloting with other languages, scaling with French

SOCIAL IMPACT

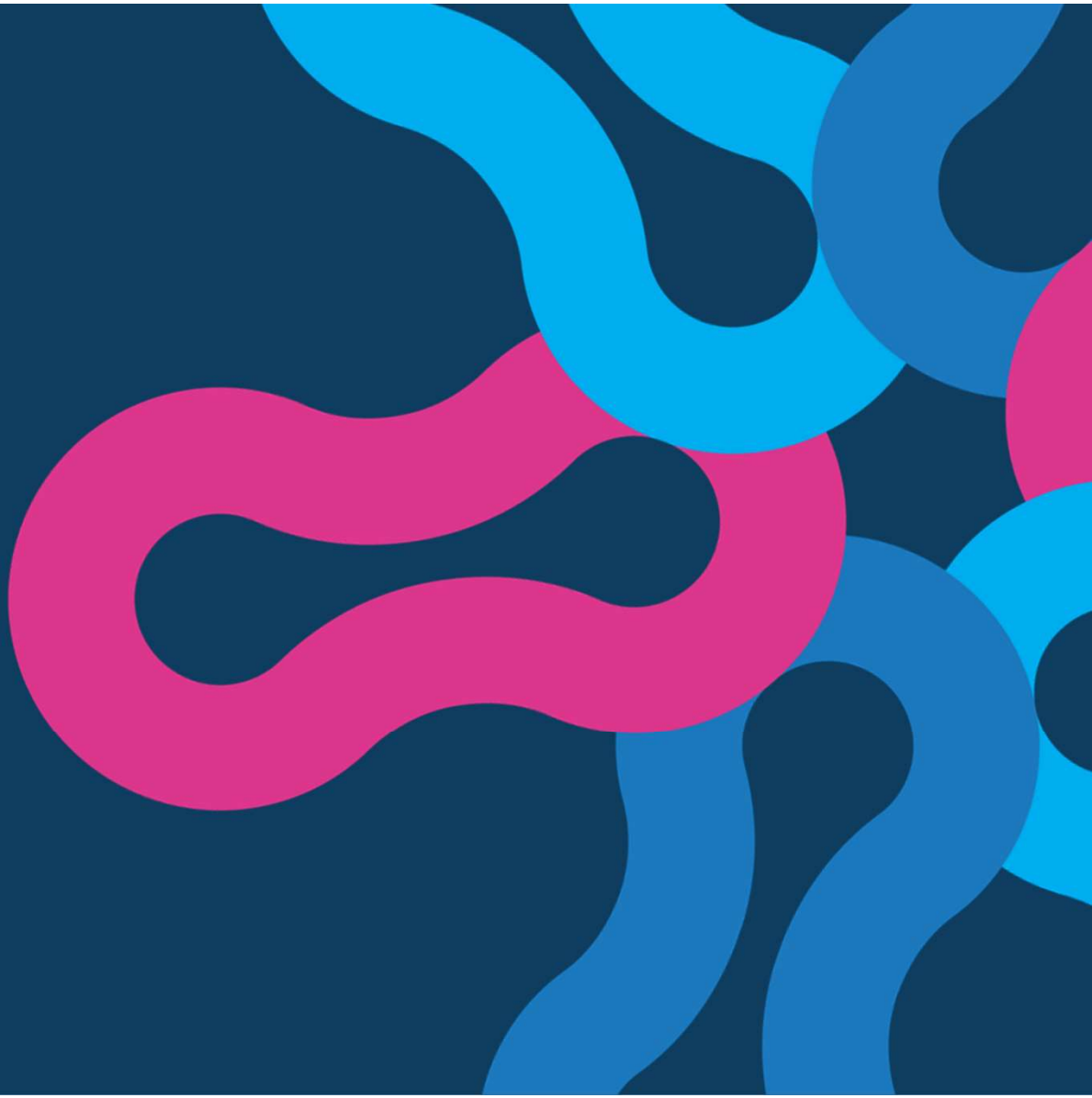
MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Transcription & Translation

Factors to Consider:

- Audio quality significantly affects output
- “Common” vs “Uncommon” languages
- Accents matter
- AI “helping” to make conversations “make sense”
- Glossary tools are (currently) minimally helpful
- Human review of AI output is essential

Qualitative Coding



Qualitative Coding

Pilot Process

MAXQDA

Claude

OpenAI

Gemini

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

AI Coding Configuration

AI coding depends on guiding the model — grounding it in the study, constraining how it applies codes, and verifying every result.

Inputs & Grounding

What the model sees

- Project context
- Code examples
- Codebook scope

Grounds the model in the specific study and codebook rather than generic priors

Coding Behavior

How the model applies codes

- Sensitivity
- Coding frequency
- Code overlap
- Quote length

Aligns the model's output with the research team's coding conventions

Validation & Quality Control

How each coding is validated

- Confidence threshold
- Quote verifier
- Multi-pass consistency
- Borderline handling

Ensures every coding passes a validation gate before it is retained

Model & Reproducibility

Which model runs the analysis, and how consistently

- Model selection
- Reasoning effort
- Codebook shuffle

Selects the appropriate model per task and controls determinism and ordering bias

Dashboard

ORGANIZATION

Knowledge Management

Literature Review

Qualitative Analysis

TOOLS

AI Chat

ADMIN

Users

Usage & Cost

Audit Log

AI Models

My usage

Sign Out

4206 coded passages in disagreement · 0 resolved · 4206 open

Raters

ref: Reference · Senegal Reform RBF - Manual Only · 9 of 9 selected

Select all | Deselect all

Reference: **Reference · Senegal Reform RBF - Manual Only · May 26, 1:34 PM**

Compared against: **All 8**

Filters

Scope: All 13 themes · all 65 codes · all 12 transcripts

Themes: **All 13**

Codes: **All 65**

Transcripts: **All 12**

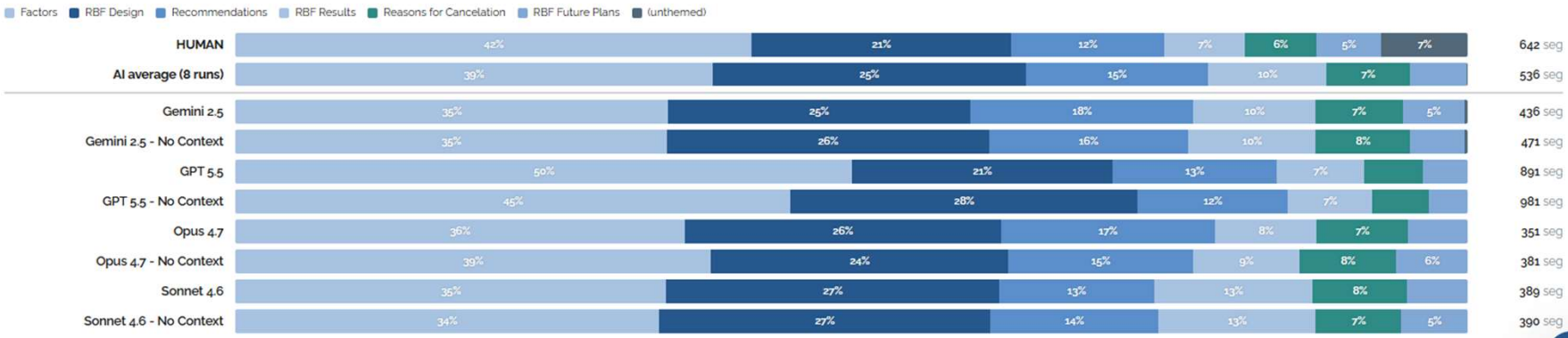
Apply filters

Visualization

Reference + 8 AI runs

VIEW: Themes | Codes | Heatmap | Segment length | Cost vs. quality | Hallucination

Each row is a rater. Bar segments show what fraction of their total coding went into each top-level theme. The "AI average" row directly below the reference is the mean share across every selected candidate. Hover a segment for the count; click a theme in the legend to highlight it across raters.



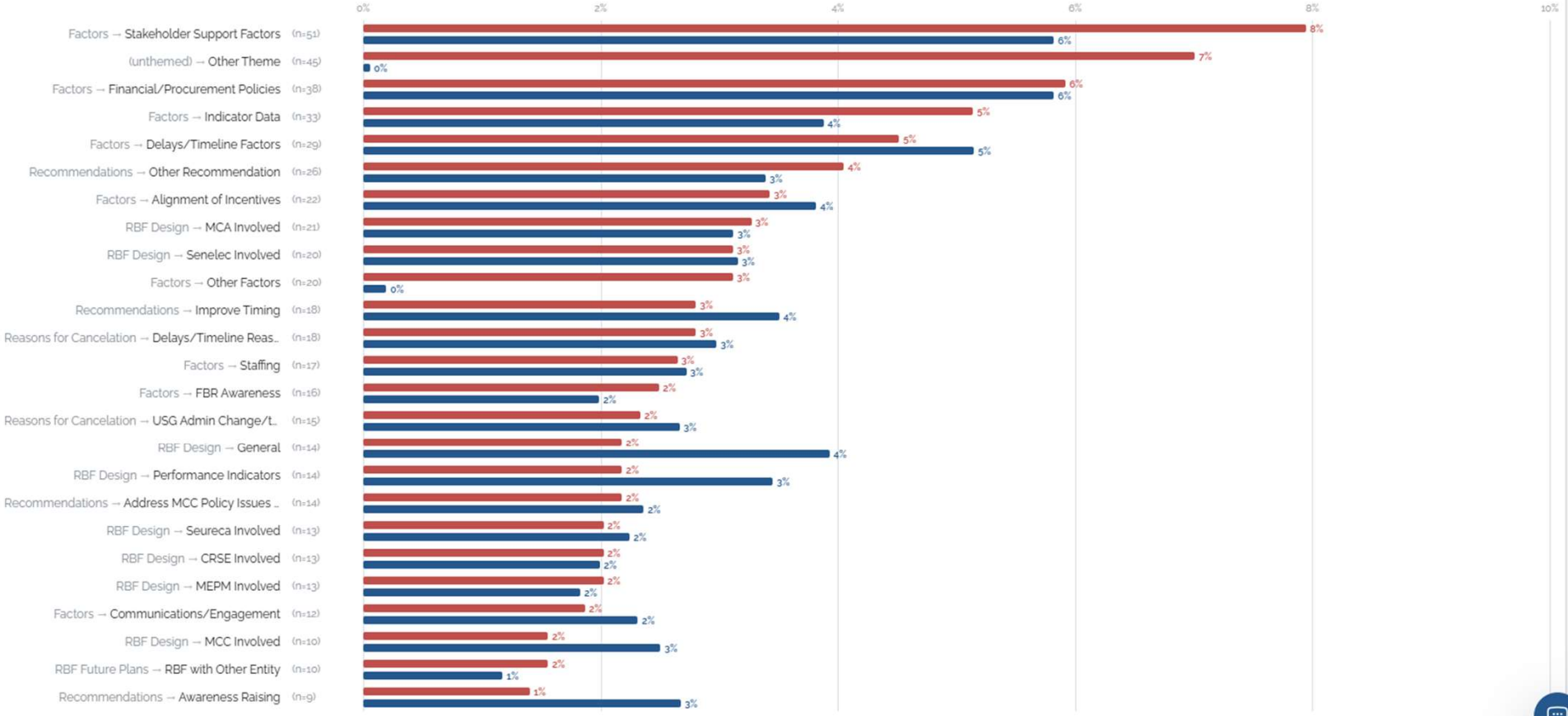
Tip — the human reference and AI rows that match the same colour profile share an analytical focus. Where AI bars lean heavily toward themes the human didn't touch (e.g., **RBF Design**), the model is coding material the researcher chose to skip for this corpus.

MCC Senegal Reform (REVISED) | 1. Inputs → 2. Refinement → 3. Coding → 4. Reconciliation → 5. Results → 6. Reporting

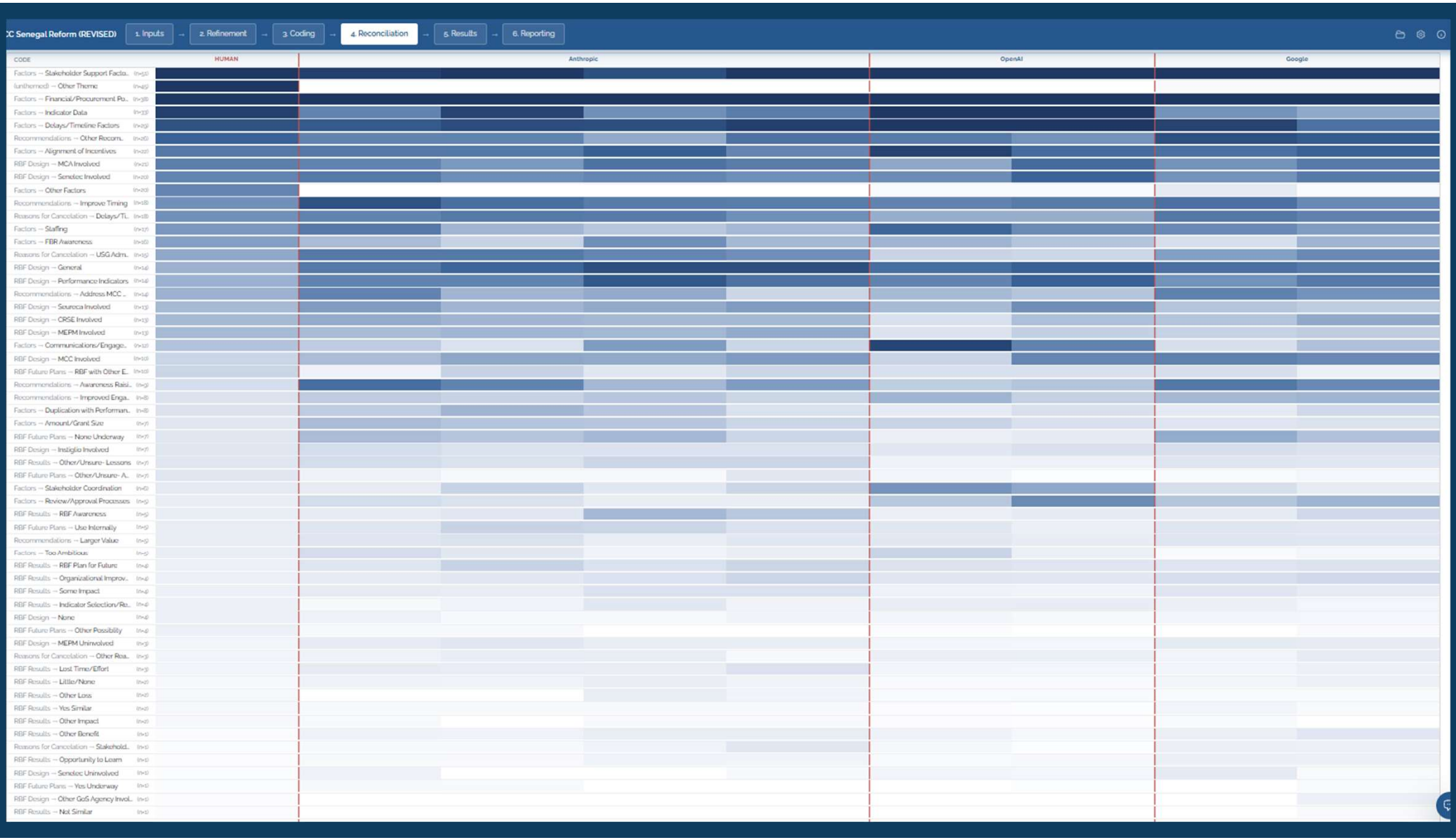
VIEW: Themes | Codes | Heatmap | Segment length | Cost vs. quality | Hallucination

Each pair of bars compares the share of the human's coding (red) vs the AI average across 8 of 8 candidate runs (blue). Codes are ranked by human-reference prevalence (descending), then by AI average for codes the human didn't use.

Sort by: Human-frequency desc | Codes: Top 25 | Theme: All themes | Rounding: On | Runs in average: 8/8



Showing 25 of 62 codes


C Senegal Reform (REVISED)
1. Inputs
2. Refinement
3. Coding
4. Reconciliation
5. Results
6. Reporting
CODE
HUMAN
Anthropic
OpenAI
Google
Factors → Stakeholder Support Facto.. (n=51)
(unthemed) → Other Theme (n=45)
Factors → Financial/Procurement Pa.. (n=38)
Factors → Indicator Data (n=33)
Factors → Delays/Timeline Factors (n=29)
Recommendations → Other Recom.. (n=26)
Factors → Alignment of Incentives (n=22)
RBF Design → MCA Involved (n=21)
RBF Design → Senelec Involved (n=20)
Factors → Other Factors (n=20)
Recommendations → Improve Timing (n=18)
Reasons for Cancelation → Delays/Ti.. (n=18)
Factors → Staffing (n=17)
Factors → FBR Awareness (n=16)
Reasons for Cancelation → USG Adm.. (n=15)
RBF Design → General (n=14)
RBF Design → Performance Indicators (n=14)
Recommendations → Address MCC .. (n=14)
RBF Design → Seureca Involved (n=13)
RBF Design → CRSE Involved (n=13)
RBF Design → MEPM Involved (n=13)
Factors → Communications/Engage.. (n=12)
RBF Design → MCC Involved (n=10)
RBF Future Plans → RBF with Other E.. (n=10)
Recommendations → Awareness Raisi.. (n=9)
Recommendations → Improved Enga.. (n=8)
Factors → Duplication with Performan.. (n=8)
Factors → Amount/Grant Size (n=7)
RBF Future Plans → None Underway (n=7)
RBF Design → Instiglio Involved (n=7)
RBF Results → Other/Unsure- Lessons (n=7)
RBF Future Plans → Other/Unsure- A.. (n=7)
Factors → Stakeholder Coordination (n=6)
Factors → Review/Approval Processes (n=5)
RBF Results → RBF Awareness (n=5)
RBF Future Plans → Use Internally (n=5)
Recommendations → Larger Value (n=5)
Factors → Too Ambitious (n=5)
RBF Results → RBF Plan for Future (n=4)
RBF Results → Organizational Improv.. (n=4)
RBF Results → Some Impact (n=4)
RBF Results → Indicator Selection/Re.. (n=4)
RBF Design → None (n=4)
RBF Future Plans → Other Possibility (n=4)
RBF Design → MEPM Uninvolved (n=3)
Reasons for Cancelation → Other Rea.. (n=3)
RBF Results → Lost Time/Effort (n=3)
RBF Results → Little/None (n=2)
RBF Results → Other Loss (n=2)
RBF Results → Yes Similar (n=2)
RBF Results → Other Impact (n=2)
RBF Results → Other Benefit (n=1)
Reasons for Cancelation → Stakehold.. (n=1)
RBF Results → Opportunity to Learn (n=1)
RBF Design → Senelec Uninvolved (n=1)
RBF Future Plans → Yes Underway (n=1)
RBF Design → Other GoS Agency Invol.. (n=1)
RBF Results → Not Similar (n=1)

MCC Senegal Reform (REVISED) | 1. Inputs → 2. Refinement → 3. Coding → 4. Reconciliation → 5. Results → 6. Reporting

Visualization | Metrics | Agreement | Compare | Finalize

4206 coded passages in disagreement · 0 resolved · **4206** open 0%

Raters ref: Reference · Senegal Reform RBF - Manual Only · 9 of 9 selected — Select all | Deselect all

Reference: **Reference · Senegal Reform RBF - Manual Only · May 26, 1:34 PM**

Compared against: **All 8**

Filters Scope: All 13 themes · all 65 codes · all 12 transcripts

Themes: **All 13** | Codes: **All 65** | Transcripts: **All 12**

Apply filters

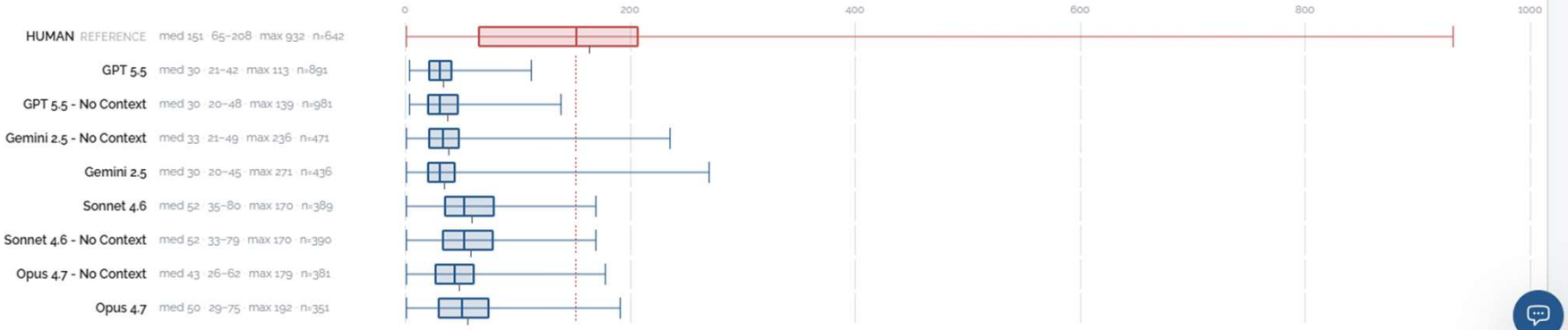
Visualization

Reference + 8 AI runs

VIEW: Themes | Codes | Heatmap | Segment length | Cost vs. quality | Hallucination

Each row is one rater's segment-length distribution. The box spans p25–p75 (the middle 50%), the vertical line inside is the median, and the whiskers reach the min and max. The black tick is the mean. A dashed red line crosses every candidate row at the human reference's median, so AI distributions shifted left or right pop visually.

Theme: All themes | Code: All codes



Raters ref: Reference · Senegal Reform RBF - Manual Only · 9 of 9 selected

Reference: **Reference · Senegal Reform RBF - Manual Only · May 26, 1:34 PM**

Compared against: **All 8**

Filters Scope: All 13 themes · all 65 codes · all 12 transcripts

Themes: **All 13**

Codes: **All 65**

Transcripts: **All 12**

Apply filters

Comparison | Inter-Rater Reliability | Segment Length

Filters

	DESCRIPTIVE						AGREEMENT							
Rater	Transcripts	Codes	Segments	Codes / Tx	Segs / Tx	Mean Seg Length	Attention F1	Theme Agreement	Code Agreement	Segment F1	Code F1	Cohen's κ	Gwet's AC1	Coverage
★ Reference · Senegal Reform RBF - Manual Only reference	12 / 12	59	642	25.9	53.5	164	—	—	—	—	—	—	—	—
Opus 4.7 - No Context Claude Opus 4.7 · No context	12 / 12	52 (−7 / 12%)	381 (−261 / 41%)	22.2 (−3.8 / 14%)	31.8 (−21.8 / 41%)	48 (−116 / 71%)	0.07	0.30 (n=47)	0.00 (n=47)	0.00	0.78	0.65	0.69	0.72
Opus 4.7 Claude Opus 4.7	12 / 12	54 (−5 / 8%)	351 (−291 / 45%)	21.8 (−4.1 / 16%)	29.3 (−24.3 / 45%)	56 (−108 / 66%)	0.09	0.37 (n=62)	0.00 (n=62)	0.00	0.78	0.66	0.70	0.72
Gemini 2.5 - No Context Gemini 2.5 Pro · No context	12 / 12	60 (+1 / 2%)	471 (−171 / 27%)	24.8 (−1.2 / 5%)	39.3 (−14.3 / 27%)	38 (−126 / 77%)	0.08	0.33 (n=67)	0.03 (n=67)	0.00	0.79	0.65	0.68	0.77
GPT 5.5 - No Context GPT-5.5 · No context	12 / 12	56 (−3 / 5%)	981 (+339 / 53%)	27.8 (+1.9 / 7%)	81.8 (+28.3 / 53%)	37 (−127 / 77%)	0.11	0.34 (n=139)	0.02 (n=139)	0.01	0.79	0.64	0.66	0.82
Sonnet 4.6 - No Context Claude Sonnet 4.6 · No context	12 / 12	54 (−5 / 8%)	390 (−252 / 39%)	23.8 (−2.2 / 8%)	32.5 (−21.0 / 39%)	58 (−106 / 65%)	0.08	0.31 (n=61)	0.00 (n=61)	0.00	0.76	0.61	0.66	0.73
Gemini 2.5 Gemini 2.5 Pro	12 / 12	57 (−2 / 3%)	436 (−206 / 32%)	23.4 (−2.5 / 10%)	36.3 (−17.2 / 32%)	35 (−129 / 79%)	0.09	0.32 (n=76)	0.04 (n=76)	0.00	0.79	0.66	0.69	0.75
Sonnet 4.6 Claude Sonnet 4.6	12 / 12	53 (−6 / 10%)	389 (−253 / 39%)	23.4 (−2.5 / 10%)	32.4 (−21.1 / 39%)	59 (−105 / 64%)	0.09	0.29 (n=76)	0.03 (n=76)	0.00	0.77	0.62	0.67	0.73
GPT 5.5 GPT-5.5	12 / 12	57 (−2 / 3%)	891 (+249 / 39%)	27.6 (+1.7 / 6%)	74.3 (+20.8 / 39%)	33 (−131 / 80%)	0.11	0.34 (n=124)	0.04 (n=124)	0.01	0.79	0.63	0.66	0.81
∑ Average across 8 runs	12.0 / 12	55.4	536.3	24.3	44.7	46	0.09	0.32	0.02	0.00	0.78	0.64	0.68	0.76

Qualitative Coding

Pilot Results:

- Coding quality is variable and depends on run specification
  - Model, codebook structure, context, prompt design, etc.
  - MaxQDA did not offer any customization
- Context improves performance

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Qualitative Coding

Pilot Results:

- Overall accuracy is improving, but variable
  - Accuracy stats only give a partial view
  - Conceptual vs word-based codes
  - False positives vs false negatives
  - Impact on findings and segments
- Having an audit trail is important (chat bots don't provide)
- Configuration matters, but no "right" configuration
- Overall, more testing and refinement is needed
  - LLM advancement also a factor

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Qualitative Coding

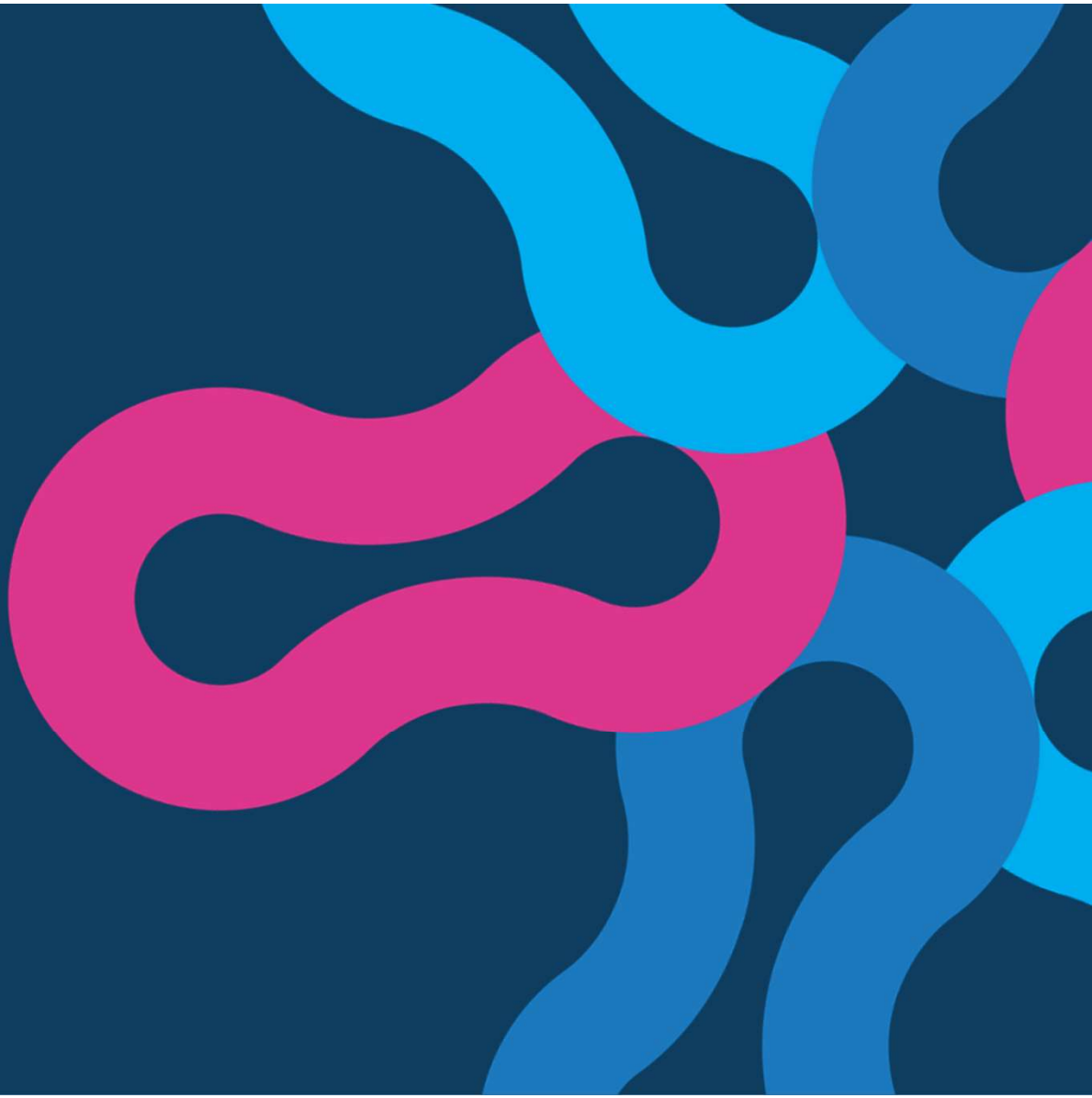
Next Steps:

- Replicate approach on other studies
  - Different sector, length, language, respondent types
- Identify configurations with stable performance across studies
- Adapt research processes for AI workflow
- Publish findings

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Closing Remarks



MCC's Operational Context

Varied understanding, appreciation and use

- Within MCC, half of staff are using AI, half are not.

Common Use	Uncommon
• Research • Summarization	• Data Analysis • Coding • Visualization • Technical Work

Benefits and Challenges

The Promise

- Analyzing large documents or datasets
- Knowledge management / finding information
- Reducing administrative burden
- Improving speed or quality of writing

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

Benefits and Challenges

The Promise	The Peril
• Analyzing large documents or datasets • Knowledge management / finding information • Reducing administrative burden • Improving speed or quality of writing	• Cognitive effects • Improper use • Accuracy • Data privacy • Environmental impact

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

MCC Questions on AI pilot proposals

Data Privacy

- How is data held?
- What are the encryption standards?
- Will indirect identifiers be included in AI-enhanced de-dentification?
- Who will have access to the AI workspace?

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

MCC Questions on AI pilot proposals

Data Privacy

Quality

- What principles will be used for crafting effective prompts?
- How will the LLM be customized?
- When and how are humans reviewing AI outputs?
- Is AI used in developing the codebook, or just in applying it to transcripts?
- How will "output quality" be determined?

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

MCC Questions on AI pilot proposals

Data Privacy

Quality

Reproducibility and Transparency

- How will prompts and replies be documented and saved?
- How will temperature controls be used to ensure more consistent outputs?

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

MCC Questions on AI pilot proposals

Data Privacy
Quality
Reproducibility and Transparency
Cost

- How will the cost compare to not using any AI?

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

MCC Questions on AI pilot proposals

Data Privacy
Quality
Reproducibility and Transparency
Cost

Learning

- What can we learn from this pilot that could be applied to future evaluations?

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

New technology, same principles

- Cost-effective use of technology
- Transparent and ethical
- Credibility is key
- Reducing research bias

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA

global
evaluation
initiative

DISCUSSION / Q&A


global
evaluation
initiative
Thank you!
www.glocalevalweek.org

Qualitative Coding

Factors to Consider:

- Structured contextual information
- Codebook development
- Prompt optimization/customization
- Desired coding results (segments/coding)
- What are you optimizing for?

SOCIAL IMPACT

MILLENNIUM CHALLENGE CORPORATION
UNITED STATES OF AMERICA