

Exercise E — Grade the output

(Leveraging) AI for Research Purposes · Class 3 · 20 October 2026

The question

Three AI tools were given exactly this question:

"What is the typical level of agreement between human coders and large language models in qualitative coding? Please give the agreement values reported in the literature (for example Cohen's kappa or percent agreement), say which study each value comes from, and end with a reference list with full references, including DOIs."

The three answers

All three were generated on **7 October 2026** and are printed **verbatim**: nothing corrected, nothing removed.

	Tool	Settings
Output 1	DeepSeek, free version	web search off : the model answers from what it learned
Output 2	Claude Sonnet	web search on
Output 3	Claude Sonnet	allowed to read only five papers and nothing else, like a notebook tool you upload sources to. The five papers are listed at the end of Output 3.

A different chatbot would make different mistakes. The method below works on all of them.

The code

Each item on your sheet is a **claim together with the source it is attributed to**. Give it one code:

Code	Meaning
A	The source does not exist. No DOI resolves to it, nothing in Crossref, PubMed, Europe PMC or Google Scholar matches. (An identifier that resolves to a <i>different, unrelated</i> paper counts as A.)
B	The source exists, but the reference is wrong: authors, year, title, journal or identifier do not match.
C	The source exists, but does not say what it is cited for: the number, the model, the task or the finding is not in it, or is different.

D	The source exists and supports the claim.
---	---

If a reference is wrong (B) **and** the claim is not in the source, use **C**: the claim is what you are grading.

For a claim with no source at all, or a generalisation across several studies, ask: *do the sources given support it?* If not, **C**.

How to check. doi.org · search.crossref.org · pubmed.ncbi.nlm.nih.gov · europepmc.org · scholar.google.com · arxiv.org. All public, no account. For **C** and **D** you must open the abstract or the paper and **find the number**. A title that looks right is not a check.

Split the work. One person per output; with four people, the fourth takes the hardest item on the sheet.

Write down, for each item: the code, and one line of evidence (what you found, where).

Report aloud (last 4 minutes): agree as a group on your seven codes and on *the most convincing error you found*. One person reads them out in the debrief.

Output 1 — DeepSeek, free, web search off

Agreement Between Human Coders and LLMs in Qualitative Coding

Overview

Reported agreement varies widely depending on task type, coding scheme complexity, model version, and whether agreement is measured per-label or aggregated. Below are values reported in the peer-reviewed literature, organized by study.

Reported Agreement Values

Xiao, Zhang & Xu (2023) — Compared GPT-3.5/GPT-4 with human coders on deductive content analysis of news texts. Reported percent agreement of approximately 87–91% for GPT-4 versus human coders on a binary/multiclass coding task, with Cohen's kappa in the 0.70–0.80 range depending on the category. GPT-3.5 performed lower, with kappa around 0.55–0.65.

Chew, Bollenbacher, Wenger, Speer & Kim (2023) — "LLM-Assisted Content Analysis" (arXiv preprint). On classifying political texts, GPT-4 achieved Cohen's kappa of approximately 0.70–0.75 against human coders, while GPT-3.5 achieved roughly 0.55–0.60. They report that GPT-4's agreement approached inter-human agreement levels for some tasks.

Gilardi, Alizadeh & Kubli (2023) — "ChatGPT outperforms crowd workers for text-annotation tasks," PNAS. ChatGPT achieved accuracy of 0.79–0.98 against expert human annotations across several annotation tasks (relevance, stance, topic, framing), often exceeding crowd-worker accuracy (0.63–0.93). They report percentage agreement rather than kappa primarily.

Törnberg (2023) — "ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages," arXiv preprint. GPT-4 achieved accuracy of 0.78–0.89 and Cohen's kappa of 0.65–0.79 across political annotation tasks, exceeding expert and crowd-worker performance in some cases.

Zhu et al. (2023) — "ChatGPT and knowledge of natural language processing," and related NLP annotation work. Reported Cohen's kappa between GPT-4 and human annotators of approximately 0.60–0.75 on NLP annotation tasks.

Deng, Zhang, Zhang, Liu & Wang (2023) — "Can ChatGPT replace human annotators?" (arXiv). Reported percent agreement of 70–85% and kappa of 0.50–0.70 across several text classification datasets, with performance varying by task difficulty.

Sanders, Ulin, Snow, Mosley & Grady (2024) — "The Paradox of Automation as Anti-Bias Intervention" / related work on LLM coding in social science. Reported Cohen's kappa of approximately 0.60–0.70 for GPT-4 on qualitative coding of interview transcripts, compared to human-human kappa of 0.70–0.85.

Meng, Zhang, Zhang & Yang (2023) — "Large language models for qualitative research" (arXiv). Reported percent agreement of 75–90% and kappa of 0.55–0.75 for GPT-4 on coding open-ended survey responses, with lower agreement on more abstract or interpretive codes.

Padiyath, Zhang, Nelson & Hovy (2024) — "The Role of LLMs in Qualitative Research" (arXiv). Reported Cohen's kappa between GPT-4 and human coders ranging from 0.45 to 0.75 depending on code specificity, with higher agreement for concrete, well-defined codes and lower for abstract or context-dependent codes.

Gebreegziabher, Zhang, Wu & Wu (2024) — "LLM-Assisted Qualitative Coding" (CHI). Reported percent agreement of 80–88% and Cohen's kappa of 0.62–0.78 between GPT-4 and human coders in a thematic analysis task, noting that agreement was highest for descriptive codes and lowest for interpretive/theoretical codes.

Bano, Zowghi, Zhang & Zhang (2024) — "LLMs for qualitative analysis in software engineering" (arXiv). Reported Cohen's kappa of 0.55–0.72 between GPT-4 and human coders on software engineering interview data.

Zhang, Zhang, Zhang & Zhang (2024) — "Can LLMs replace human coders?" (arXiv). Reported percent agreement of 72–89% and kappa of 0.50–0.75 across multiple qualitative datasets, with substantial variation by code type.

General Patterns

1. Model generation matters: GPT-4 typically achieves kappa ~0.60–0.80; GPT-3.5 typically ~0.45–0.65.
2. Task type matters: Concrete, deductive, well-specified codes yield higher agreement (kappa up to ~0.80); abstract, inductive, interpretive codes yield lower agreement (kappa ~0.40–0.60).
3. Human-human baseline: Human-human kappa in qualitative coding is typically ~0.70–0.85, so LLMs often approach but rarely exceed this ceiling.
4. Percent agreement tends to be higher (~75–90%) than kappa because it does not correct for chance.
5. Few studies report full confusion matrices or per-code breakdowns, which limits comparability.

Caveats

- Many of the above are preprints and have not undergone full peer review.
- Agreement values are not directly comparable across studies due to differences in codebooks, datasets, number of categories, and whether agreement is computed per-code or overall.

- Some studies report accuracy against a human "gold standard" rather than inter-coder agreement, which is a different metric.

Reference List

1. Xiao, Z., Zhang, Y., & Xu, Y. (2023). Can ChatGPT replace human coders? A study on deductive content analysis. arXiv preprint arXiv:2304.11771 . <https://doi.org/10.48550/arXiv.2304.11771>
2. Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). LLM-assisted content analysis: Using large language models to support deductive coding. arXiv preprint arXiv:2306.14924 . <https://doi.org/10.48550/arXiv.2306.14924>
3. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120. <https://doi.org/10.1073/pnas.2305016120>
4. Törnberg, P. (2023). ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. arXiv preprint arXiv:2304.06588 . <https://doi.org/10.48550/arXiv.2304.06588>
5. Zhu, Y., Zhang, Y., Zhang, Y., & Yang, Y. (2023). ChatGPT and knowledge of natural language processing. arXiv preprint arXiv:2303.07151 . <https://doi.org/10.48550/arXiv.2303.07151>
6. Deng, Y., Zhang, Y., Zhang, Y., Liu, Y., & Wang, Y. (2023). Can ChatGPT replace human annotators? arXiv preprint arXiv:2304.10889 . <https://doi.org/10.48550/arXiv.2304.10889>
7. Sanders, N. E., Ulin, L., Snow, B., Mosley, A., & Grady, S. (2024). The paradox of automation as anti-bias intervention. *Cardozo Law Review*, forthcoming. (Related LLM coding analyses reported in supplementary materials.)
8. Meng, Y., Zhang, Y., Zhang, Y., & Yang, Y. (2023). Large language models for qualitative research. arXiv preprint arXiv:2305.12345 . <https://doi.org/10.48550/arXiv.2305.12345>
9. Padiyath, A., Zhang, Y., Nelson, L., & Hovy, D. (2024). The role of LLMs in qualitative research. arXiv preprint arXiv:2402.12345 . <https://doi.org/10.48550/arXiv.2402.12345>
10. Gebreegziabher, S. A., Zhang, Y., Wu, Z., & Wu, Q. (2024). LLM-assisted qualitative coding. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3613904.3642345>
11. Bano, M., Zowghi, D., Zhang, Y., & Zhang, Y. (2024). LLMs for qualitative analysis in software engineering. arXiv preprint arXiv:2403.12345 . <https://doi.org/10.48550/arXiv.2403.12345>
12. Zhang, Y., Zhang, Y., Zhang, Y., & Zhang, Y. (2024). Can LLMs replace human coders? arXiv preprint arXiv:2404.12345 . <https://doi.org/10.48550/arXiv.2404.12345>

Note: Several DOIs and arXiv identifiers above are illustrative/approximate based on the cited works' typical locations. Before citing in formal work, please verify each DOI and arXiv ID directly against the publisher or arXiv record, as preprint identifiers and page numbers may have changed upon peer-reviewed publication.

Output 2 — Claude Sonnet, web search on

Agreement between human coders and LLMs in qualitative coding — literature summary

Short answer: Reported agreement spans a wide range (fair to almost-perfect), roughly $\kappa \approx 0.3$ – 0.9 and percent agreement ≈ 60 – 95% , depending on task type (deductive vs. inductive), codebook complexity, model, and prompting strategy. The most common range for well-specified deductive coding with a clear codebook is **substantial agreement ($\kappa \approx 0.6$ – 0.8)**, comparable to typical human-human intercoder reliability. Agreement drops toward "moderate" or "fair" ($\kappa \approx 0.3$ – 0.5) for inductive coding, culturally/emotionally loaded themes, or ambiguous codebooks.

Specific values by study

- **Bijker, Merkouris, Dowling & Rodda (2024)**, *JMIR* — ChatGPT content analysis of forum posts on sugar-reduction experiences. Intercoder κ ranged **0.72–0.82** across inductive coding schemes and **0.58–0.73** across unconstrained coding/structured coding-matrix schemes. Precision for detecting change mechanisms: 66–88%.
- **Liu, Zambrano, Baker, Barany, Ocumpaugh, Zhang, Pankiewicz, Nasir & Wei (2025)**, *Journal of Learning Analytics* ("Qualitative Coding with GPT-4: Where it Works Better") — kappa between GPT-4 and human coders ranged **0.72–0.95** depending on the code/dataset; performance varied strongly by code type (some codes near-perfect, others weak).
- **Borse, Chatta Subramaniam & Rebello (2025)**, *PERC Proceedings* (arXiv:2508.14764) — ChatGPT-4o/4.5 vs. human raters coding audio transcripts of physics problem-solving discussions. After prompt/hyperparameter optimization: **substantial agreement ($\kappa > 0.6$) for three themes, moderate agreement for one.**
- **Sakaguchi, Sakama & Watari (2025)**, *JMIR* — ChatGPT-4 vs. human researchers, thematic analysis of 30 Japanese clinical interviews. Agreement **>80%** for descriptive/concrete themes (e.g., "personal experience of a sacred moment," "building relationships"), but only **~30%** for themes requiring deep cultural/emotional interpretation (e.g., "fate").
- **General pattern reported across multiple comparative studies** (aggregated in search of the literature): among studies reporting quantitative agreement indices, **median percent agreement ≈ 0.80 (range 0.19–1.00)**, **median Cohen's kappa ≈ 0.74 (range 0.38–0.82)**, median cosine-similarity coefficient ≈ 0.58 (range 0.47–0.80).
- **Multi-LLM thematic analysis study** (arXiv:2512.20352, "Multi-LLM Thematic Analysis with Dual Reliability Metrics") — Gemini $\kappa = 0.907$ (cosine 95.3%), GPT-4o $\kappa = 0.853$ (cosine 92.6%), Claude $\kappa = 0.842$ (cosine 92.1%) vs. human coders — all "almost perfect" by conventional κ bands.
- **Fleiss' kappa studies** (multiple-rater designs): Fleiss' κ between human coders and GPT-4 ranged **0.44–0.50** (moderate), while three independent GPT-4 runs agreed with each other at Fleiss' $\kappa \approx 0.87$ (near-perfect) — i.e., the model can be more internally consistent with itself than with humans.
- **Systematic-review screening context** (LLM-assisted title/abstract screening, not thematic coding per se): Cohen's $\kappa = 0.41$ under standard inclusion criteria (moderate), 0.38 under lenient criteria (fair) — lower than typical thematic-coding agreement, reflecting a different (binary inclusion) task.
- Other studies surveyed report **fair-to-moderate** agreement with κ as low as **0.332–0.403**, and GPT-4 reaching κ of **0.81, 1.00, 0.79** for specific individual codes within a larger, more mixed set.

Standard interpretation scale used across these papers (Landis & Koch convention)

$\kappa < 0.40$ = poor/fair; $0.40\text{--}0.60$ = moderate; $0.60\text{--}0.80$ = substantial; $\kappa > 0.80$ = almost perfect/excellent.

Factors that move agreement up or down (reported in the literature)

- Agreement is higher for **deductive coding with a well-specified codebook** than for **inductive/open coding**.
- Providing the model a **rationale/justification requirement** improved agreement (mean κ 0.68 vs. 0.59 without it).
- Presenting **one code at a time** rather than the whole codebook in a single prompt improved agreement (mean κ 0.68 vs. 0.60).
- Agreement is lower for themes requiring **cultural, emotional, or implicit interpretation** than for concrete/descriptive themes.
- Different LLMs (Gemini, GPT-4o, Claude) show measurably different agreement levels on the same task.

Reference list

1. Bijker R, Merkouris SS, Dowling NA, Rodda SN. ChatGPT for Automated Qualitative Research: Content Analysis. *J Med Internet Res*. 2024;26:e59050. <https://doi.org/10.2196/59050>
2. Liu X, Zambrano AF, Baker RS, Barany A, Ocumpaugh J, Zhang J, Pankiewicz M, Nasir N, Wei Z. Qualitative Coding with GPT-4: Where it Works Better. *Journal of Learning Analytics*. 2025;12(1):169–185. <https://doi.org/10.18608/jla.2025.8575>
3. Borse NS, Chatta Subramaniam R, Rebello NS. Investigation of the Inter-Rater Reliability between Large Language Models and Human Raters in Qualitative Analysis. *2025 Physics Education Research Conference (PERC) Proceedings*. Preprint: arXiv:2508.14764. <https://doi.org/10.48550/arXiv.2508.14764>
4. Sakaguchi K, Sakama R, Watari T. Evaluating ChatGPT in Qualitative Thematic Analysis With Human Researchers in the Japanese Clinical Context and Its Cultural Interpretation Challenges: Comparative Qualitative Study. *J Med Internet Res*. 2025;27:e71521. <https://doi.org/10.2196/71521> (PMC12062757)
5. Multi-LLM Thematic Analysis with Dual Reliability Metrics: Combining Cohen's Kappa and Semantic Similarity for Qualitative Research Validation. Preprint: arXiv:2512.20352. <https://doi.org/10.48550/arXiv.2512.20352>

Note: items 3 and 5 are preprints (arXiv) at the time of this search; DOIs given are the arXiv DOIs. If the researcher needs only peer-reviewed, published sources, items 1, 2, and 4 are journal-published (JMIR, Journal of Learning Analytics) with standard DOIs; item 3 is a published conference-proceedings paper (PERC 2025) but I could only confirm the arXiv DOI, not a separate PER-Central/AAPT DOI — worth checking per-central.org directly if a non-arXiv DOI is required. I was not able to independently verify a DOI for the aggregated "median $\kappa = 0.74$ / median % agreement = 0.80" synthesis figure — it surfaced from search-engine synthesis across multiple comparative studies rather than one identifiable paper, so I'd treat that specific aggregate statistic as indicative rather than citable to one source; flag this to the researcher if they want to use that number directly.

Output 3 — Claude Sonnet, restricted to five papers

Agreement between human coders and LLMs in qualitative coding: values reported across five studies

Across these five papers, reported agreement is highly task- and metric-dependent, but a rough pattern emerges: **percent agreement is usually high (70–99%)**, while **chance-corrected statistics (Cohen's κ , Fleiss κ , Krippendorff's α) vary much more widely**, from near-zero/poor to near-perfect, often depending on code prevalence and task type (deductive vs. inductive coding).

1. Gilardi, Alizadeh & Kubli (2023) — ChatGPT vs. crowd workers and trained annotators (text annotation, not strictly qualitative coding but closely related)

- Intercoder agreement (percentage of same label by two annotators), averaged across four datasets and multiple tasks (relevance, stance, topic, frame detection):
- MTurk crowd workers: ~56%
- Trained human annotators: ~79%
- ChatGPT (temperature = 1): ~91%
- ChatGPT (temperature = 0.2): ~97%
- ChatGPT's zero-shot accuracy (agreement with trained-annotator gold standard) exceeded MTurk by about 25 percentage points on average; accuracy for binary relevance tasks ranged from 59% to 83% depending on dataset.

2. Leas et al. (2024) — ChatGPT vs. human annotators for adverse-event detection in Reddit posts

- Any adverse event (AE) detection: 94.4% agreement (9436/10,000), prevalence- and bias-adjusted Fleiss κ = 0.95
- Serious AE detection: 99.3% agreement (9931/10,000), κ = 0.96 (abstract) / the results text also reports 99.3% (9331/10,000... note: table shows 9331, abstract says 9931 — reported as in source) with κ = 0.96
- By MedWatch subcategory, agreement ranged from 99.4% (κ = 0.98, "other" serious outcomes) to 99.9% (κ = 0.99) for specifically defined outcomes (e.g., hospitalization, life-threatening).
- For comparison, the paper cites a prior specialized AE-detection software benchmark at 94.5% agreement, κ = 0.89.

3. Balt et al. (2025) — LLAMA3 (few-shot) vs. human researchers coding psychosocial-autopsy interviews

- Binary classification task (does a code appear in a segment): mean accuracy = 0.84 (SD 0.09); Cohen's κ = 0.68 ("substantial agreement" per Landis & Koch criteria), though performance varied widely by code (κ ranging roughly from 0.24 to 1.00 across individual codes, per their Table 2).
- Sliding-window segment-identification task: accuracy = 0.67 (no κ reported for this task).
- Summarization task (rated by two human researchers as good/adequate/poor): about 80% of summaries rated "adequate" or "good" (61%/20%/19% good/adequate/poor by rater 1; 56%/24%/21% by rater 2).

4. van den Berkmortel, Westra, Gifford & van de Baan (2026) — GPT-5 and Gemini 2.5 Pro vs. human coding of a Dutch newspaper corpus (nurse sentiment, COVID-19)

- Human–human baseline (double coding, ATLAS.ti intercoder tool): Krippendorff's α = 0.703 ("acceptable" agreement).
- Human vs. GPT-5 (deductive coding of predefined codebook): Krippendorff's α = 0.487

- Human vs. Gemini 2.5 Pro: Krippendorff's $\alpha = 0.507$
- Both AI models "overcoded" relative to humans (181 and 183 total code applications vs. 138 for humans), which the authors argue drove the lower α values — reflecting differing "quotation unitization" rather than purely disagreement on meaning.
- Note: this paper cites Li et al. (not one of the five source papers here) as reporting "only moderate agreement (Cohen's $\kappa = 0.401$)" between GPT-4 and human researchers on patient-interview coding — this is a secondary citation within the paper, not independently verified from Li et al.'s own text.

5. Hill et al. (2026) — ChatGPT-5, Claude 4 Sonnet, QualiGPT vs. blinded human analysts (healthcare focus-group thematic analysis)

Deductive coding (against expert-consensus adjudicated reference standard):

- Human blinded analysts (mean): 92.7% agreement (95% CI 91.6–93.9); Cohen's $\kappa = 0.34$ (95% CI 0.26–0.41); Gwet's AC1 = 0.92 (95% CI 0.90–0.93)
- LLMs (mean of ChatGPT-5, Claude 4 Sonnet, QualiGPT): 93.5% agreement (95% CI 92.6–94.5); Cohen's $\kappa = 0.34$ (95% CI 0.26–0.40); Gwet's AC1 = 0.93 (95% CI 0.92–0.94)
- Individual models: ChatGPT-5 — 93.5% agreement, $\kappa = 0.31$, AC1 = 0.93; Claude 4 Sonnet — 93.8% agreement, $\kappa = 0.35$, AC1 = 0.93; QualiGPT — 93.3% agreement, $\kappa = 0.34$, AC1 = 0.93
- Pre-adjudication human-human agreement (CH vs. GS, before third-reviewer adjudication): 99.0% (95% CI 98.4–99.4); $\kappa = 0.931$ (95% CI 0.891–0.966); Gwet's AC1 = 0.990 (95% CI 0.984–0.994)
- Thematic-level Cohen's κ (by individual code, human and LLM vs. reference panel) ranged from –0.05 to 0.54 across the 10 codes — illustrating substantial topic-by-topic variability even within one study.
- All LLMs were statistically non-inferior to human analysts ($p < 0.0001$); ChatGPT-5 ($p = 0.048$) and Claude 4 Sonnet ($p = 0.039$) were statistically superior to humans on this deductive task.

Inductive (open-ended theme generation): measured via a 5-point Likert agreement scale against the reference panel rather than κ — Human Analyst 1 mean 3.643, Human Analyst 2 mean 4.071, ChatGPT-5 mean 4.071, Claude 4 Sonnet mean 3.643, QualiGPT mean 3.929. Only ChatGPT-5 reached the pre-specified non-inferiority margin.

Hallucination/error rates (inductive analysis, quote verification): strict hallucination rate 1.2% (SD 2.1%); expanded hallucination rate 8.6% (SD 5.1%); comprehensive error rate 12.4% (SD 5.1%).

Summary take

Percent agreement is consistently high across all five studies (generally 70–99%), but this is not a reliable summary on its own: chance-corrected metrics tell a more mixed story. Fleiss κ for a clearly defined, low-ambiguity task (AE detection: Leas et al.) reached 0.95–0.99 — near-perfect. Cohen's κ for genuinely interpretive qualitative coding tasks (Balt et al.; Hill et al.) landed in the "moderate" to "substantial" range (0.34–0.68), and Krippendorff's α for deductive code application in van den Berkmortel et al. was markedly lower (0.49–0.51) than the human-human baseline (0.70), driven by LLM overcoding. Hill et al. explicitly note that κ is strongly attenuated by low code prevalence, which is why they also report Gwet's AC1 (consistently ~0.92–0.93 for both humans and LLMs) as a prevalence-robust complement.

References

1. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*. 2023;120(30):e2305016120. <https://doi.org/10.1073/pnas.2305016120>
2. Leas EC, Ayers JW, Desai N, Dredze M, Hogarth M, Smith DM. Using Large Language Models to Support Content Analysis: A Case Study of ChatGPT for Adverse Event Detection. *Journal of Medical Internet Research*. 2024;26:e52499. <https://doi.org/10.2196/52499>
3. Balt E, Salmi S, Bhulai S, Vrinzen S, Eikelenboom M, Gilissen R, Creemers D, Popma A, Mérelle S. Deductively coding psychosocial autopsy interview data using a few-shot learning large language model. *Frontiers in Public Health*. 2025;13:1512537. <https://doi.org/10.3389/fpubh.2025.1512537>
4. van den Berkmortel B, Westra D, Gifford R, van de Baan FC. Analyzing Nurse Sentiment in a Dutch Regional Newspaper During the COVID-19 Pandemic: Comparative Content Analysis of GPT-5, Gemini 2.5 Pro, and Human Coding. *Journal of Medical Internet Research*. 2026;28:v28i2e98374. <https://doi.org/10.2196/98374>
5. Hill C, Dahil A, Simpson G, Hardisty D, Keast J, Pinn CK, Dambha-Miller H. Large language models for thematic analysis in healthcare research: A blinded mixed-methods comparison with human analysts. *PLOS Digital Health*. 2026;5(4):e0001189. <https://doi.org/10.1371/journal.pdig.0001189>

Group sheets

One per group. Seven items each: three from Output 1, two from Output 2, two from Output 3.

How to use your sheet. Find your group's sheet below (Turing, Lovelace, Hopper or Shannon). Each item is one claim taken from one of the three answers, with the number of the reference it cites in that answer's own reference list.

- **Split.** One person per output: three items from Output 1, two from Output 2, two from Output 3. With four people, the fourth takes the hardest item.
- **Find the source.** Look up the reference in the answer's list, then check it: doi.org, Crossref, PubMed, Europe PMC, Google Scholar, arXiv. Items with no reference (a summary claim) are checked against the five papers of Output 3.
- **Find the claim.** Open the abstract or the paper and look for the number, the model, the task. Words in bold tell you exactly what to check.
- **Code it.** Give each item one code, A, B, C or D (see The code, above), and write one line of evidence: what you found, and where.
- **Report.** 27 minutes for checking. In the last 4, agree on your seven codes and on the most convincing error you found: one person reads them out in the debrief, then we look at the key together.

Group Turing

Output 1

- 1. Xiao, Zhang & Xu (2023): GPT-4 agreement 87–91%, κ 0.70–0.80 on news texts. Reference 1.
- 2. Gilardi, Alizadeh & Kubli (2023): accuracy 0.79–0.98 against **expert** annotations; crowd workers 0.63–0.93. Reference 3.

- 3. Zhang, Zhang, Zhang & Zhang (2024): agreement 72–89%, κ 0.50–0.75. Reference 12.

Output 2

- 4. Bijker et al. (2024): κ 0.72–0.82 and 0.58–0.73, presented as agreement **between ChatGPT and human coders**. Reference 1.
- 5. Borse et al. (2025): $\kappa > 0.6$ for three themes, moderate for one. Reference 3.

Output 3

- 6. The summary claim: "percent agreement is usually high (70–99%)" / "consistently high across all five studies". Check it against the five papers.
- 7. Balt et al. (2025): mean accuracy 0.84 (SD 0.09), Cohen's κ 0.68, "substantial agreement". Reference 3.

Group Lovelace

Output 1

- 1. Chew et al. (2023): GPT-4 κ 0.70–0.75, GPT-3.5 κ 0.55–0.60, on political texts. Reference 2.
- 2. Zhu et al. (2023): κ 0.60–0.75 on NLP annotation tasks. Reference 5.
- 3. Meng, Zhang, Zhang & Yang (2023): agreement 75–90%, κ 0.55–0.75 on open-ended survey responses. Reference 8.

Output 2

- 4. Multi-LLM Thematic Analysis (arXiv:2512.20352): Gemini κ 0.907, GPT-4o 0.853, Claude 0.842 **"vs. human coders"**. Reference 5.
- 5. Sakaguchi et al. (2025): >80% agreement for descriptive themes, ~30% for themes such as "fate"; 30 interviews. Reference 4.

Output 3

- 6. van den Berkmortel et al. (2026): the lower α reflects **"quotation unitization"** rather than disagreement on meaning. Reference 4.
- 7. Hill et al. (2026): LLMs 93.5% agreement (95% CI 92.6–94.5), κ 0.34 (95% CI 0.26–0.40), Gwet's AC1 0.93. Reference 5.

Group Hopper

Output 1

- 1. Törnberg (2023): GPT-4 accuracy 0.78–0.89, κ 0.65–0.79 across political annotation tasks. Reference 4.
- 2. Deng et al. (2023): agreement 70–85%, κ 0.50–0.70 across several datasets. Reference 6.
- 3. Sanders, Ulin, Snow, Mosley & Grady (2024): GPT-4 κ 0.60–0.70 on interview transcripts. Reference 7.

Output 2

- 4. Liu et al. (2025): κ between GPT-4 and human coders **0.72–0.95** "depending on the code/dataset". Reference 2.
- 5. Bijker et al. (2024): κ 0.72–0.82 and 0.58–0.73, presented as agreement **between ChatGPT and human coders**. Reference 1.

Output 3

- 6. Hill et al. (2026): "which is **why** they also report Gwet's AC1 as a prevalence-robust complement". Is this the paper's reason? Reference 5.
- 7. The summary claim: Cohen's κ for Balt and Hill is "**moderate to substantial (0.34–0.68)**".

Group Shannon

Output 1

- 1. Bano, Zowghi, Zhang & Zhang (2024): κ 0.55–0.72 on software-engineering interview data. Reference 11.
- 2. Padiyath, Zhang, Nelson & Hovy (2024): κ 0.45–0.75 depending on code specificity. Reference 9.
- 3. Gebreegziabher, Zhang, Wu & Wu (2024): agreement 80–88%, κ 0.62–0.78 in thematic analysis, CHI. Reference 10.

Output 2

- 4. Multi-LLM Thematic Analysis (arXiv:2512.20352): Gemini κ 0.907, GPT-4o 0.853, Claude 0.842 "**vs. human coders**". Reference 5.
- 5. Liu et al. (2025): κ between GPT-4 and human coders **0.72–0.95** "depending on the code/dataset". Reference 2.

Output 3

- 6. Hill et al. (2026): thematic-level κ by code ranged **from –0.05 to 0.54**. Check the text **and** the table.
- 7. Leas et al. (2024): serious adverse events 99.3% — and "*the results text also reports ... 9331/10,000*". Where does 9331 appear? Reference 2.