When Does Human Input Matter? Evaluating LLM-Assisted Analysis of Public Participation in Housing Approval Process

Wenzheng Li, Julia Harten, Christos Thrampoulidis

The University of British Columbia

Funded by the New Frontiers in Research Fund (NFRF) and the Canada Postdoctoral Research Award (CPRA), SSHRC

Background

- **Computational text analysis has made large-scale public input tractable** (Cai 2021; Sanchez et al. 2023)
  - e.g., topic identification and sentiment analysis read plans (Brinkley & Stahmer 2021; Fu et al. 2023), identify policies (Wang et al. 2026) and assess public attitudes (Kim et al. 2021).

- **Conventional NLP vs. LLMs**
  - Conventional NLP needs task-specific training and struggles with nuanced meaning (Fu et al. 2023; Lore et al. 2024).
  - LLMs can use pretrained knowledge and prompts to interpret text, and outperform NLP on planning tasks (Fu et al. 2024; Rong et al. 2025).

- **Human-in-the-loop**
  - The human works as the model's collaborator, validating and refining its output at each stage (Fu et al. 2024; Zheng et al. 2025).
  - **To the planner, the process is still a black box:** one cannot tell when a result is credible, how much human input it needed, or how to balance time input and model performance.

Research objectives

Q1 Where does human input yield the highest marginal return?

One prompt grows in four cumulative stages, each adding to the one before: category names (D0), definitions (D1), annotated examples (D2), refinement on a dev. set (D3). Each stage measures marginal gain in accuracy (macro F1) per hour of human work.

Q2 How do gains vary across tasks, categories (and models)?

Three tasks, scored the same way. Inside each task, we separate categories that require inference and judgement (reasoning-based) from categories identified by an explicit cue in the text (mention-based).

Q3 What is left after the last step, and why?

We audit every error the final prompt still makes: which ones can be explained and fixed (in the prompt, the model, or the gold), and which are genuinely hard cases? We then test what predicts an error.

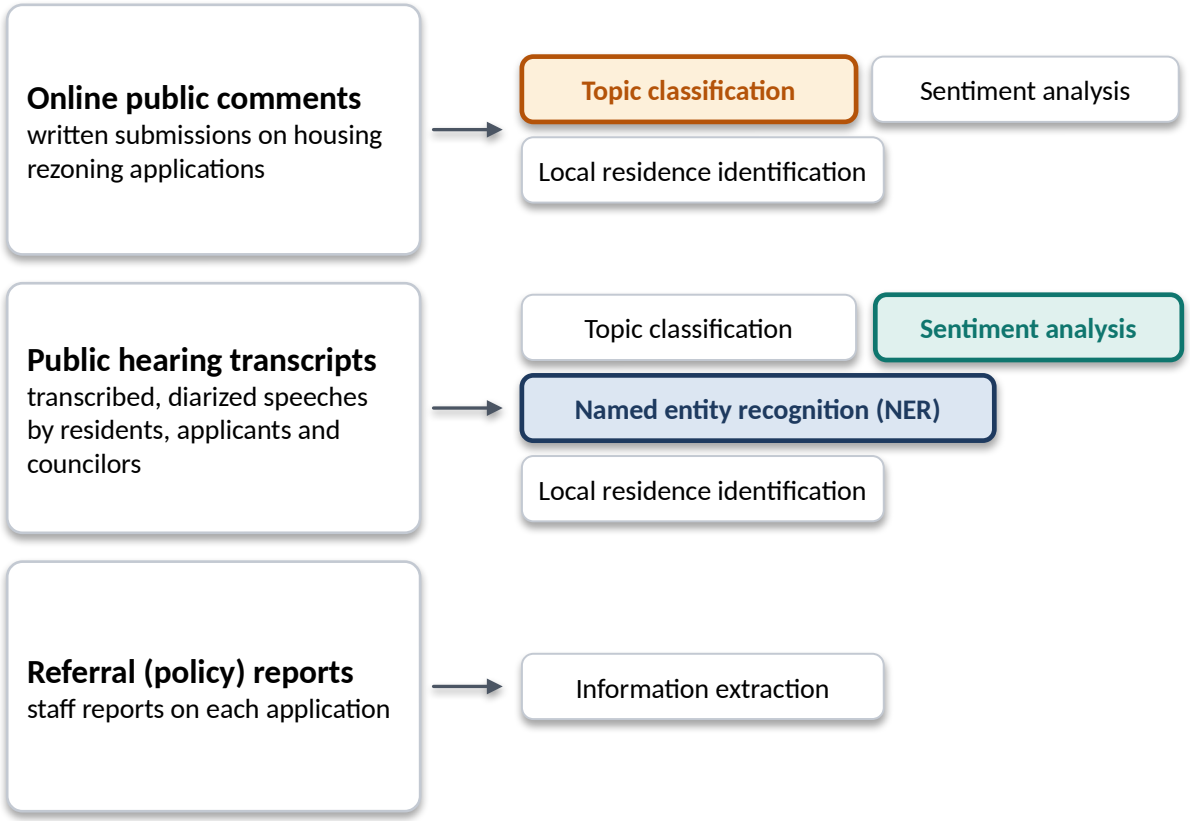
Contribution

- **A framework for understanding human–LLM collaboration**: explaining how different forms of human input, task demands, and (model capabilities) shape performance gains and the errors that remain.
- **A practical guide** for researchers and planners: where human hours pay when LLMs are used on different text tasks, and when to stop.

Data and tasks

350 housing rezoning projects in Vancouver from 2018 to 2025, collected from the City of Vancouver: Home

Public records of the housing project → Tasks



Topic classification

- Which of 14 topics (e.g., affordability, traffic) a comment raises (multi-label)
- 7 mention-based and 7 reasoning-based topics.
- Gold: 5,214 comment × topic pairs from 1,447 comments.

Sentiment analysis

- A speaker's overall stance toward the application;
- 4 classes:
  - support, oppose - mention-based
  - neutral, mixed - reasoning-based
- Gold: 198 speakers

Named entity recognition (NER)

- Who is speaking?
- 2 classes:
  - role (public speaker/councilor) - mention-based
  - name (open set) - reasoning-based
- Gold: 106 reachable names in the transcript.

Why these three:

- Cover both things LLMs do with text, classifying and extracting.
- **Each contains categories an explicit cue settles (mention-based) and categories that need judgement (reasoning-based).**

Research framework

1 Human input (D0-D3), added stage by stage to the same prompt

each stage keeps everything from the stage before - one prompt per stage

Input to D0–D2: a small sample of raw texts, read by the researcher (unlabeled); the basis for the category names, definitions and examples

D0 · Label names

The researcher names the topics/categories.

Prompt: task + topics/category names

D1 · Definitions

Writes a definition and a scope rule for each category (the codebook).

Prompt: + definitions

D2 · Examples

Selects and annotates positive and negative examples.

Prompt: + Match/Do-not-match examples

D3 · Refinement ↻ iterate

Labels dev. set(s), reviews the model's errors, rewrites rules and boundary cases; repeats.

Prompt: + rules, boundary cases, revised definitions

2 The LLM labels every unit with the stage's prompt

open models: Gemma-3-12B; scored on held-out gold

3 Measure at every stage

Accuracy—gold macro F1 at each stage, with bootstrap intervals

Human effort—hours of each stage, derived from the work log

Return on human time—ΔF1 per hour of each stage: ***across tasks***, and ***within tasks***

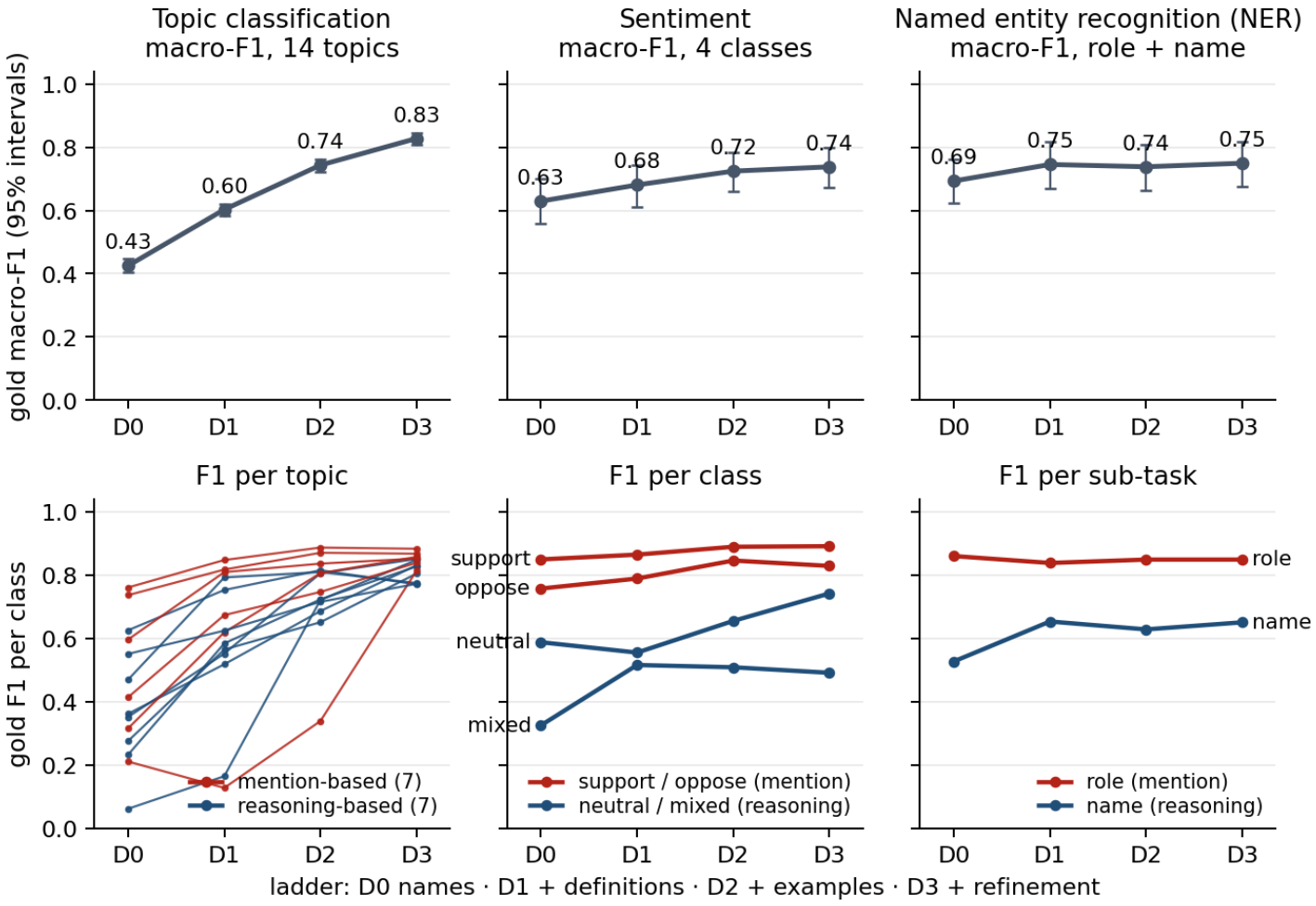
the errors left at D3

4 Error analysis of what remains after D3

Audit of every wrong answer: cause (model/prompt/ambiguity/gold), mechanism

Regression: which units stay wrong, and why, in terms of category type, text structure, and judgement cues

Results 1 - Accuracy along the ladder

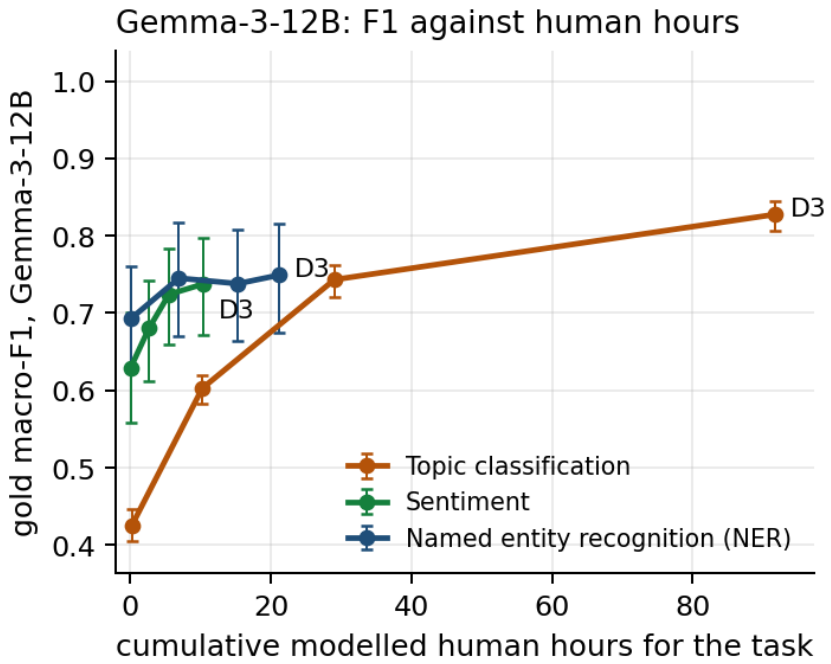


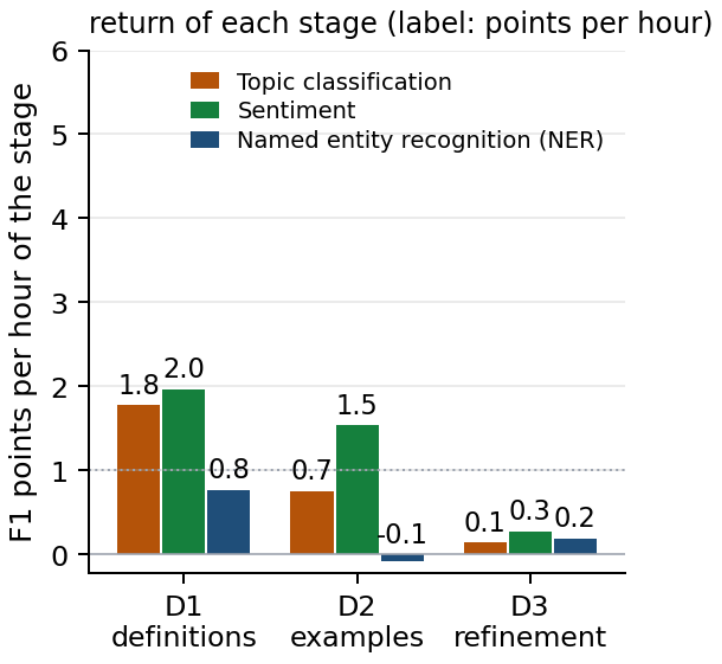
Top: gold macro-F1 of Gemma-3-12B at each stage of the ladder (bars = 95% bootstrap intervals).
Bottom: gold F1 per category (blue = reasoning-based, red = mention-based).

- Across tasks, human input raises accuracy (F1) in every task, but by very different amounts. From D0 to D3:
  - +0.40 on topic classification;
  - +0.11 on sentiment;
  - +0.06 on NER.

- Within each task,
  - the reasoning-based categories start lower and move most.
  - the mention-based ones are relatively high from the start with less improvement.

Results 2 - Return per hour, across tasks





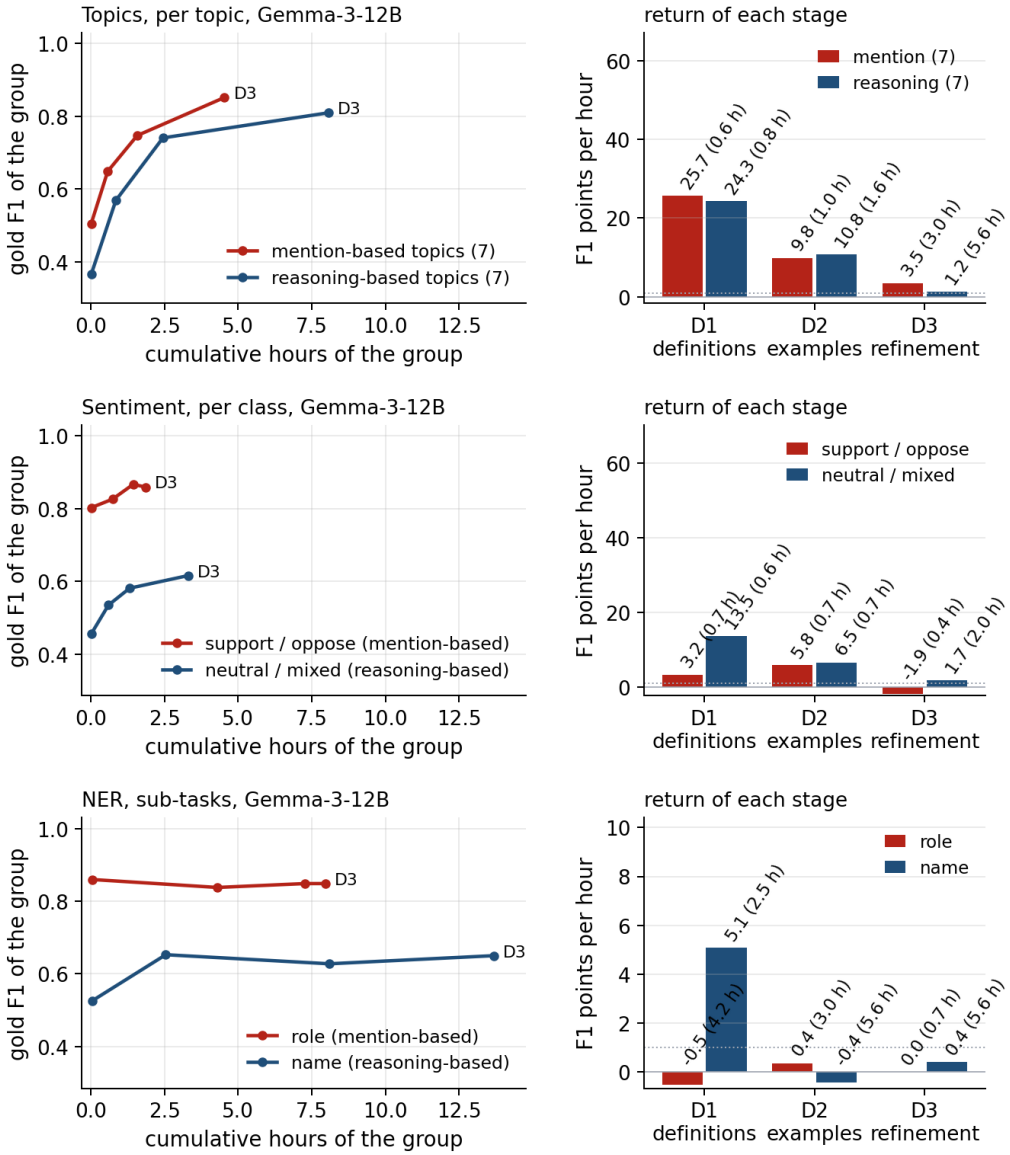
Left: *gold macro-F1 against cumulative human hours, D0 to D3.* ***Right:*** *F1 points per hour of each stage (1 point = 0.01 F1).*

- Definitions are the best buy in every task: 0.8-2 points per hour.
- Examples pay for classification (topics, sentiment), not for NER.
- Refinement costs the most and returns the least: below 1 point per hour everywhere.

Note: *Not in the curves - a fixed familiarization cost before D0, reading a small sample of raw texts to determine the topics/classes*

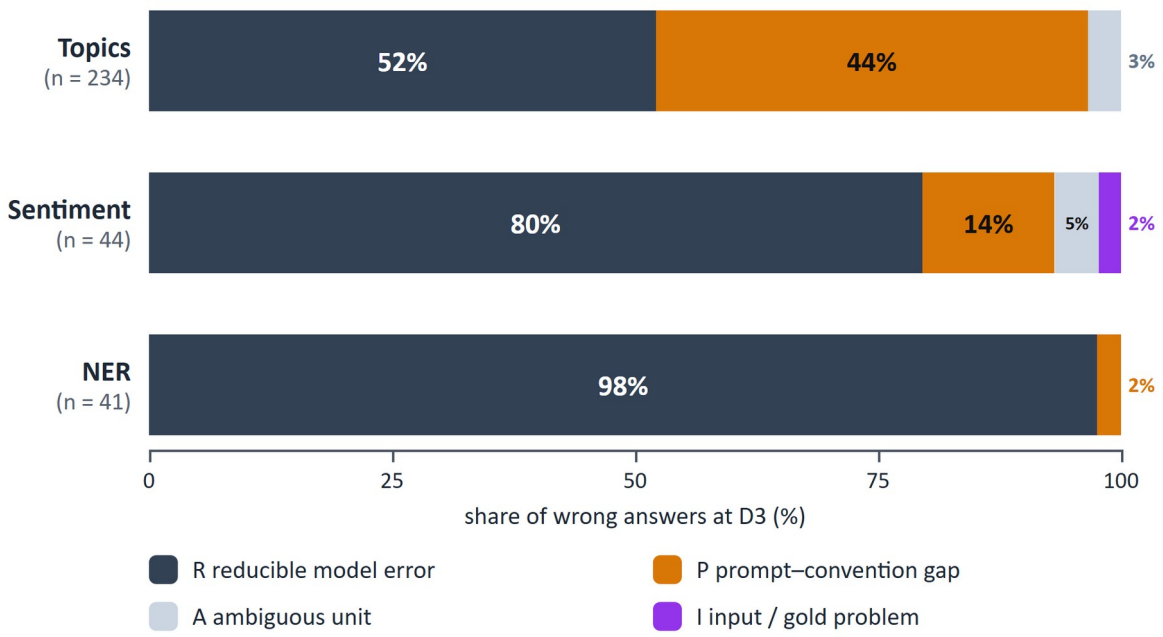
Results 3 - Return per hour, within tasks

- **More hours, lower ceiling.** Reasoning-based categories absorb more of the hours and still end lower than mention-based.
- **More room for human input.** The reasoning-based groups gain the most from D0 to D3.
- **Stage by stage.** Definitions (D1) pay for every group; after that, examples and refinement pay only where judgement is needed.



Left: *gold F1 of each category group against the hours apportioned to it, D0 to D3.* ***Right:*** *F1 points per hour of each stage, by group (1 point = 0.01 F1).*

Error analysis 1 - What is still wrong at D3



Typology of errors:

- **R reducible model error** - the model missed a rule the prompt already gave *e.g. praise without a stated position: the prompt says neutral, the model said support*
- **P prompt–convention gap** - the prompt itself allowed the wrong label *e.g. "build as much density as possible" labelled affordability under the prompt's wording*
- **A ambiguous unit** — boundary case: either label works *e.g. "28 storeys may be right … we just don't know yet": Densification, yes or no*
- **I Input/gold problem** — wrong gold label

Error analysis 2 · What explains the remaining errors?

	Topic classification		Sentiment		Named entity recognition (NER)	
	(1) structure	(2) + cues	(3) structure	(4) + cues	(5) structure	(6) + cues, name
Reasoning-based category (= 1)	0.26 (0.23)	0.11 (0.17)	2.36*** (0.44)	1.46*** (0.51)	2.03*** (0.29)	
Category prevalence in the gold (z)	0.64*** (0.09)	0.56*** (0.08)				
Label crowding: topics per comment (z)	-0.07 (0.15)	-0.40** (0.15)				
Text length: log words (z)	0.52*** (0.17)	0.55*** (0.13)	0.43** (0.20)	-0.05 (0.21)	-0.35* (0.20)	0.05 (0.21)
Councillor (ref. public speaker)			-0.48 (0.41)	-0.98* (0.50)	2.01*** (0.49)	1.03** (0.43)
Lexical cue against the gold label (z)		1.38*** (0.13)				
Decision path, label-blind raters (1–5)				1.06*** (0.22)		
Explicit stance phrase (= 1)				0.30 (0.51)		
Name outside the block (= 1)						1.07*** (0.37)
Log distance of the name (turns)						0.16 (0.28)
Observations (wrong answers)	5,214 (234)	5,214 (234)	198 (44)	198 (44)	218 (52)	106 (41)
Pseudo R^2	0.19	0.31	0.18	0.33	0.23	0.11

Logit model; Standard errors in parentheses: (1)–(2) clustered by topic (14 clusters, t13); (3)–(4) robust; (5) clustered by block;
*** p < 0.01, ** p < 0.05, * p < 0.10.

Summary

Human input: practical guidelines

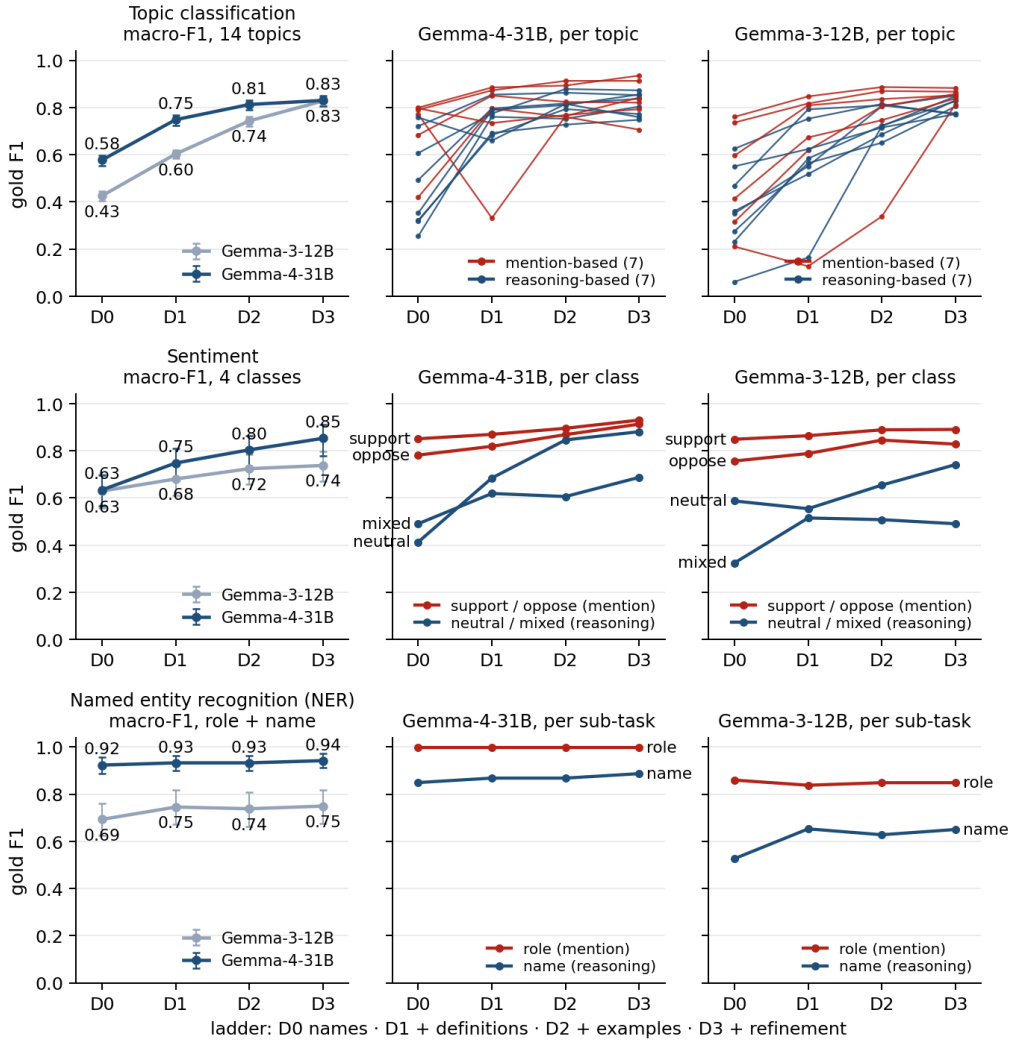
- **1 Plan by category.** Sort categories into mention-based and reasoning-based; stop early on the first group.
- **2 Always write the definitions.** The codebook stage has the best marginal benefits in every task (0.8 - 2 F1 points per hour) and is cheap (3 - 10 hours).
- 3 **Add examples for classification tasks rather than for extraction** (examples paid on topics and sentiment, nothing on NER).
- **4 Refine only where judgement is needed.** Refinement is the expensive stage, but it is the step that finally improves the performance of your tasks. Its gains usually sit on the reasoning-based categories.
- **5 Expect a ceiling.** The final prompt still leaves 4–39% of units wrong; after D3 the next hour is better spent on the audit or trying better models.

Error analysis: what is left

- **Most of the remaining errors are deterministic, not random.**
  - The great majority come from the model's own limitations (R) or from an unclear prompt (P); genuinely ambiguous cases are few.

Next step

the same ladder on Gemma-4-31B (dark) against Gemma-3-12B (light); middle and right columns per category



Next step: stronger (reasoning) models. Two questions:

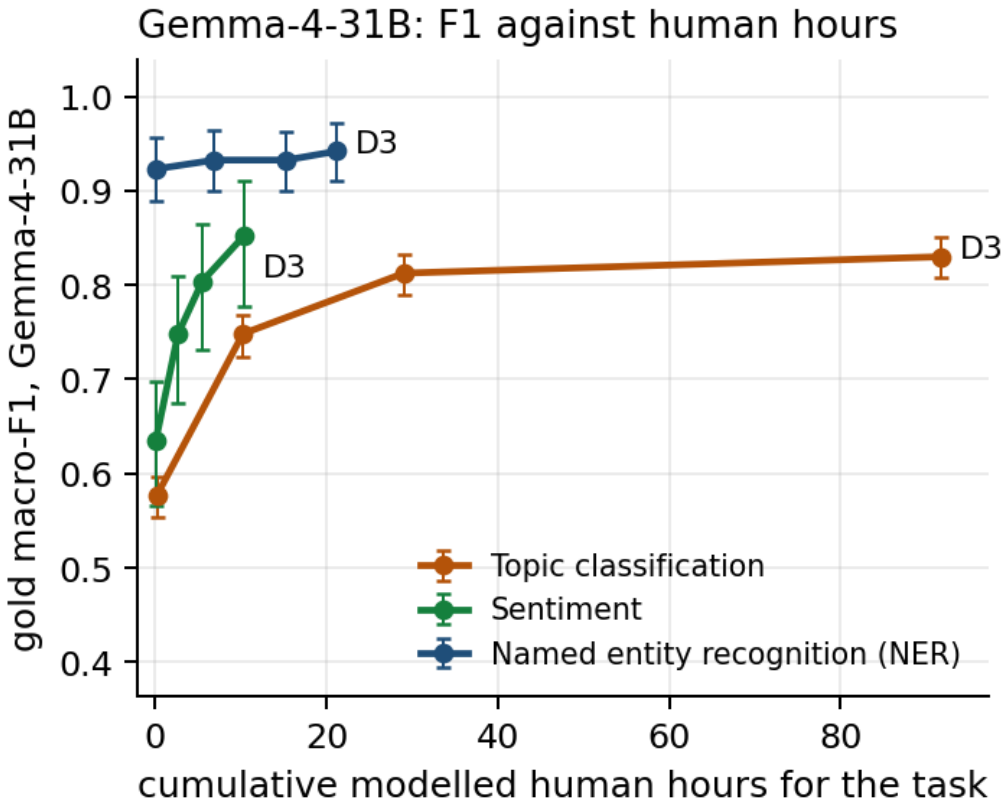
- Does the mention/reasoning gap narrow?
- How much can refinement still add?
  - mention-based categories reach peak performance with little or no refinement;
  - For reasoning-based categories, the remaining work is to reduce ambiguity in the prompt, so human time falls sharply.

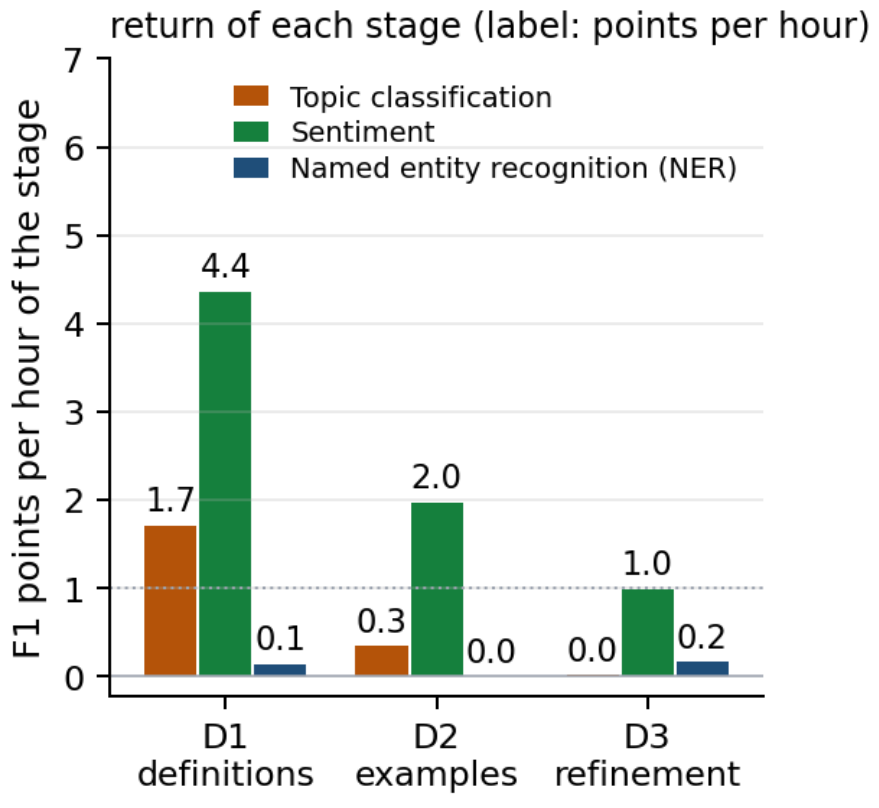
Thank you — questions and comments

Wenzheng Li (wenzheng.li@ubc.ca)

This research is funded by the Canada Postdoctoral Research Fellowship (CPRA) from the Social Sciences and Humanities Research Council of Canada (SSHRC).

Return per hour across tasks, Gemma-4-31B

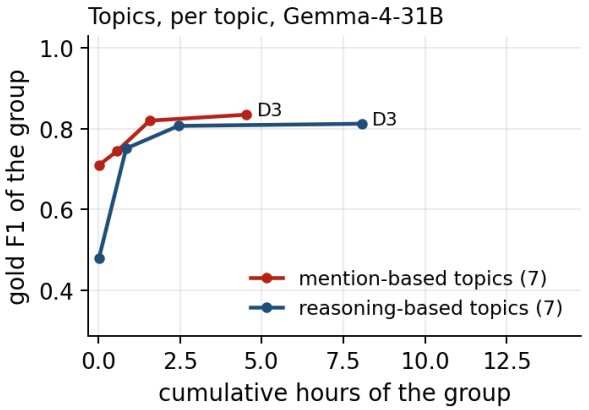


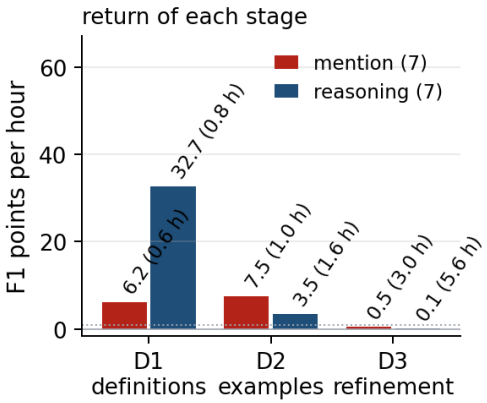


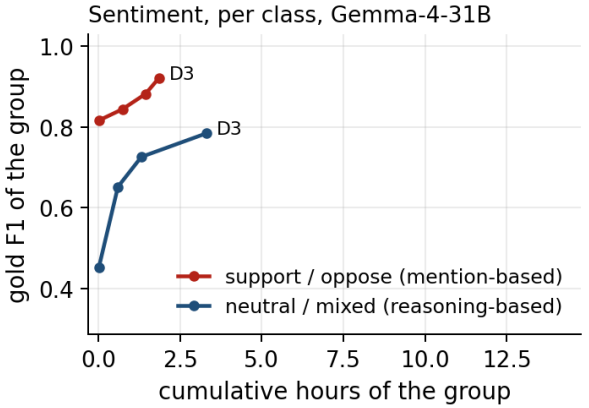
Left: *gold macro-F1 of Gemma-4-31B against cumulative human hours, D0 to D3 (the same hours as for Gemma-3-12B).*
Right: *F1 points per hour of each stage (1 point = 0.01 F1).*

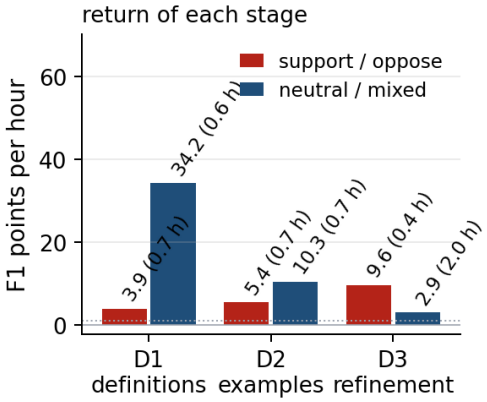
Return per hour within tasks, Gemma-4-31B

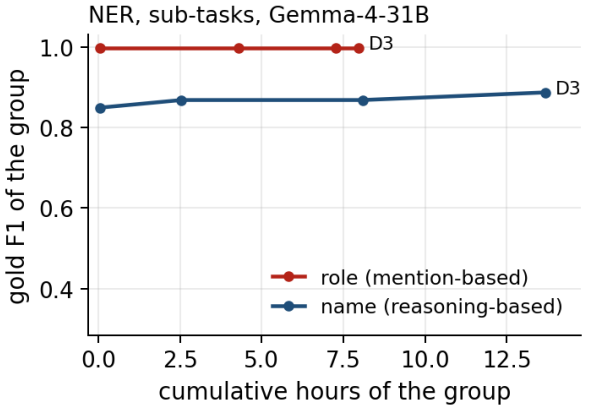
Left: *gold F1 of each category group against the hours apportioned to it, D0 to D3.*
Right: *F1 points per hour of each stage, by group (1 point = 0.01 F1). Gemma-4-31B; same hours as for Gemma-3-12B.*

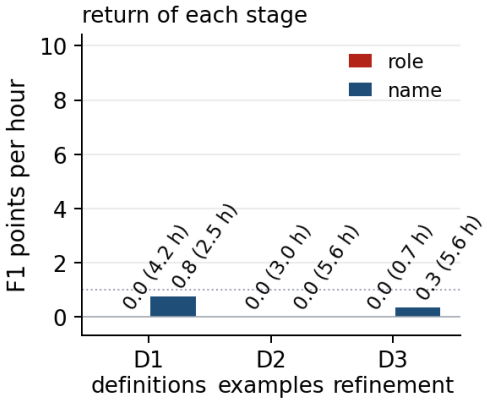




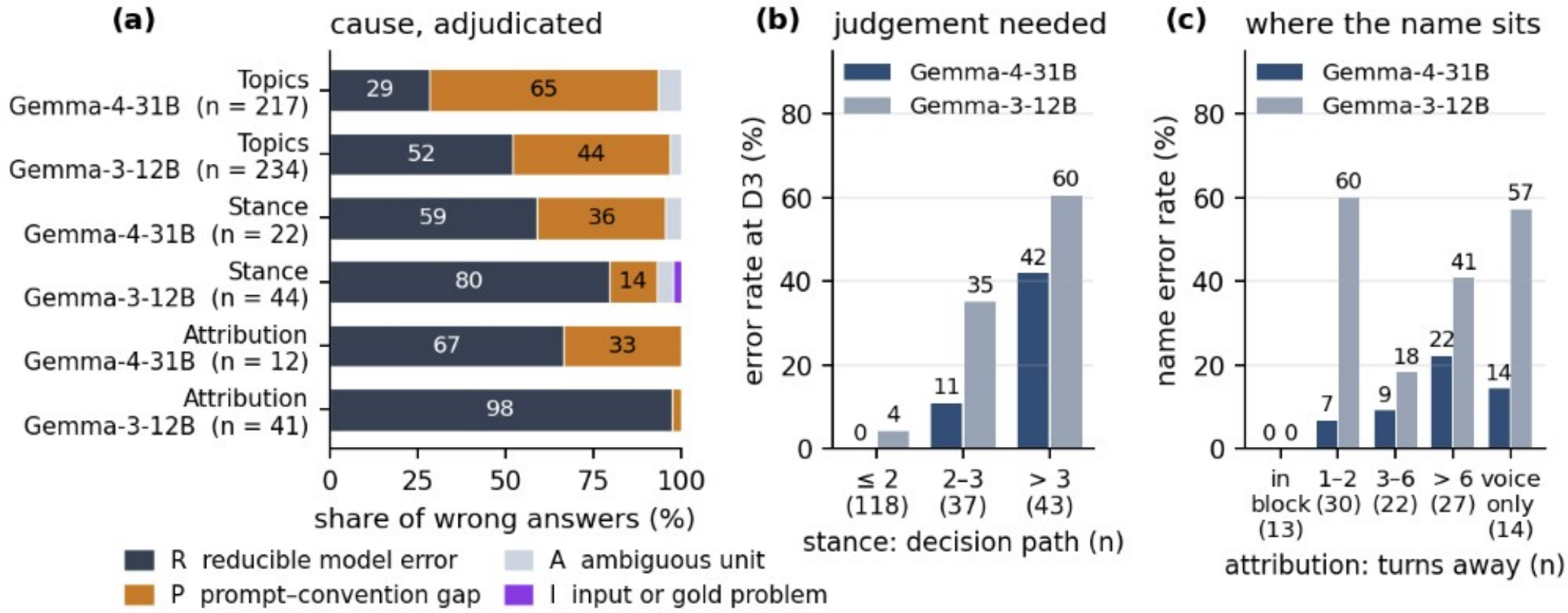








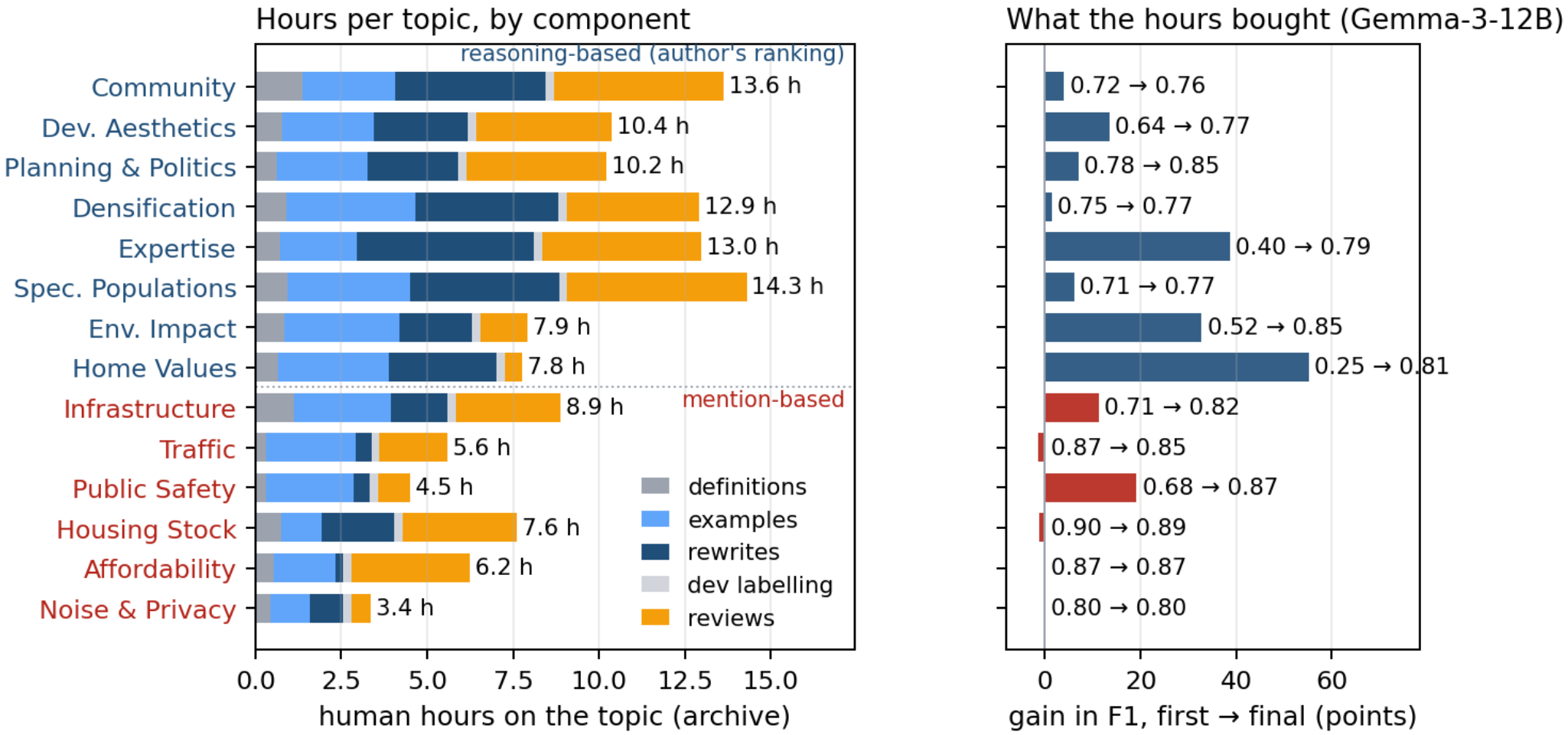
Error analysis with the stronger model (Gemma-4-31B)



(a) cause of each wrong answer, by model; (b) sentiment error rate at D3 by the raters' decision path; (c) name error rate at D3 by where the name sits

- **With Gemma-4-31B the residual shifts from the model to the prompt**: on topics, reducible errors fall from 52% to 29% of wrong answers and prompt–convention gaps rise from 44% to 65%.
- **The hard units stay hard for both models:** sentiment errors rise with the decision path (4 → 35 → 60% for Gemma 3; 0 → 11 → 42% for Gemma 4), and NER fails where the name sits outside the block or only in the voice hint.

Where the hours went on topics



Left: *human hours per topic by component (based on working log of students);*
right: *gain in gold F1 from the first to the final prompt (based on Gemma-3-12B)*

The comparison of human versus LLM or the comparison of human versus human: https://www.gov.uk/government/news/ground-breaking-use-of-ai-saves-taxpayers-money-and-delivers-greater-government-efficiency