Can LLMs Advance Occupational Coding? Evidence and Methodological Insights

Olga Kononykhina,
Ludwig Maximilian University of Munich,
Munich Center for Machine Learning

4 June 2026, UCL

From Manual to LLM coding

Manual coding: 2+ coders per occupation. Gold standard but slow, costly, inconsistent.

2Mio occupations is 10 years of FTE for a professional coder.

Rule-based coding: Uses job title text matching (e.g. CASCOT). Fast, codes into multiple taxonomies but sensitive to input.

In some cases confidently codes 20-30% of answers only

ML-based coding: Learns from labeled data (e.g. OccuCoDe). Fast, scalable, needs a lot of training data.

Accuracy around ~30-60%, 70% was reported

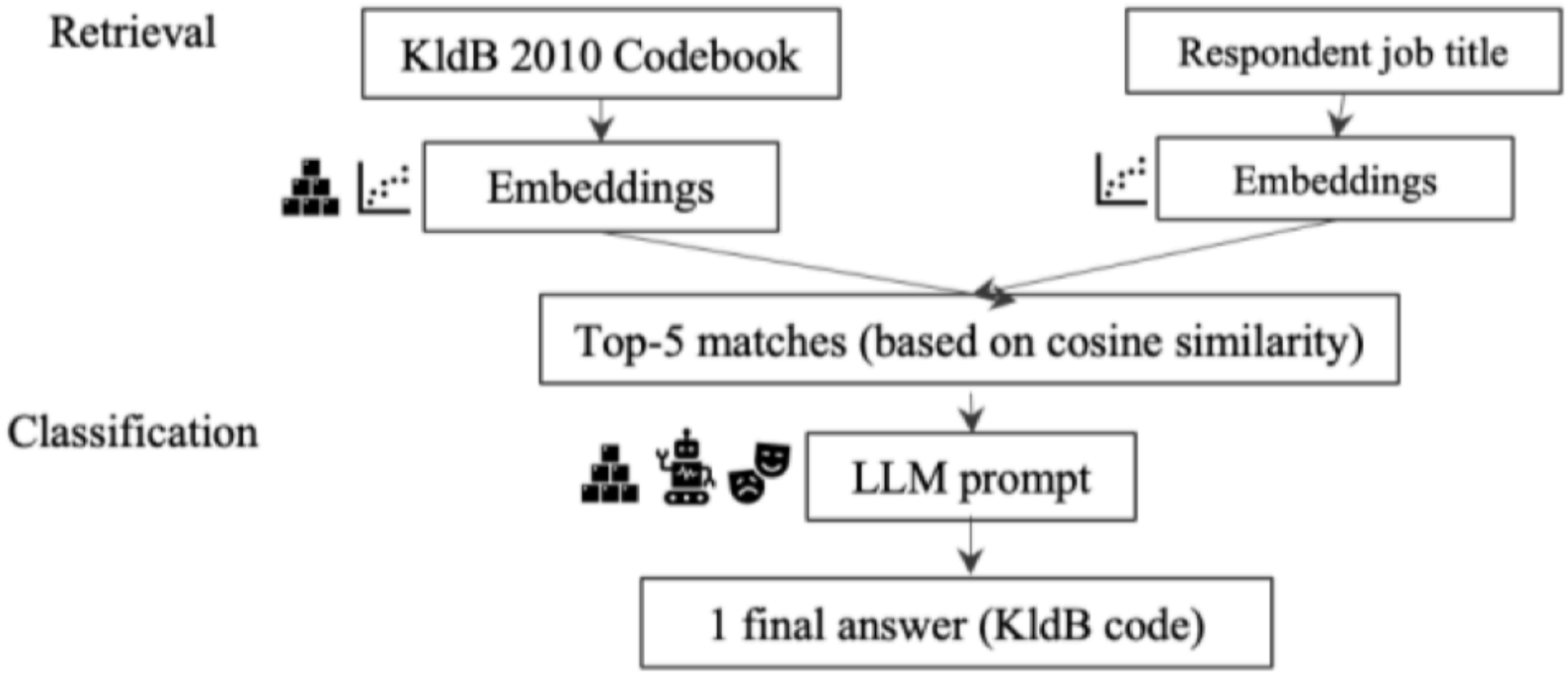
BERT (Transformers): Contextual NLP. Handles complex text, needs large training data.

Accuracy around 43-821%

LLM for occupational coding

- **Fine tuning**
  Computationally expensive
  Needs training data
  Codes into one system only

- **Retrieval or Retrieval + Ranking**
  Computationally cheaper
  Easier to adapt to different coding systems
  Easier to try newer systems

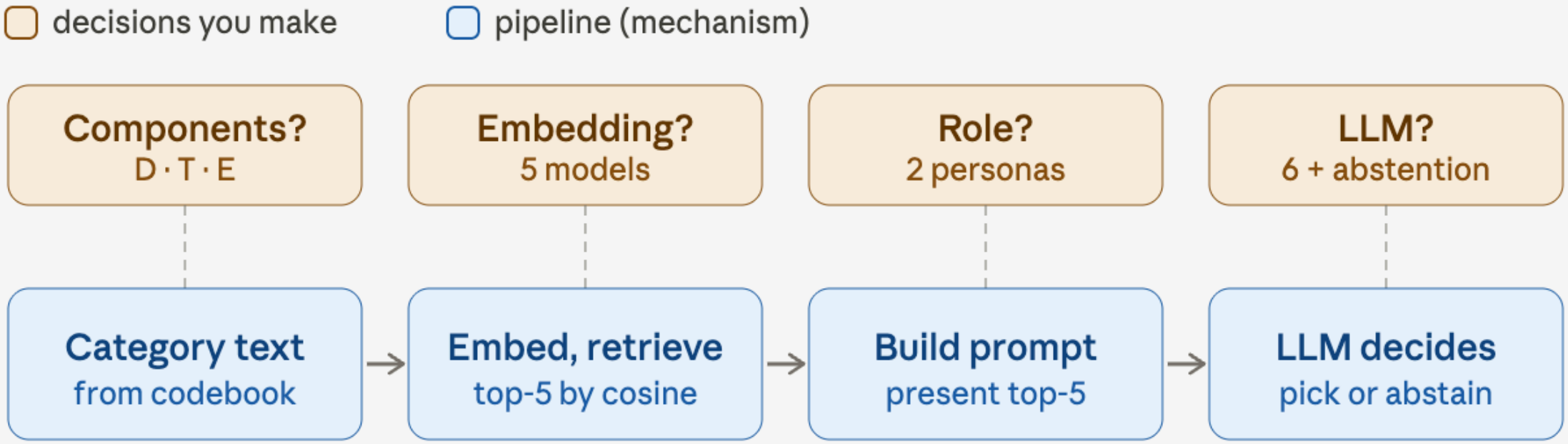

Retrieval
KldB 2010 Codebook
Respondent job title
Embeddings
Embeddings
Top-5 matches (based on cosine similarity)
Classification
LLM prompt
1 final answer (KldB code)

LLM based occupational coding

Approach	Reference	Data	Result
embeddings + classic ML classifier	OPERAS (Langezaal et al. 2023)	Job title, sector	ISCO-88/ 66%
Pure embedding similarity (bi-encoder, zero-shot or lightly tuned)	CareerBERT (Rosenberger et al. 2025); JobHop (Johary et al. 2025); MELO benchmark (Retyk et al. 2024)	job titiles, descriptions, skills	MRR@100 ~ 0.5-0.7 Top1 (ESCO)=78%
Retrieval-augmented prompting (vector search → LLM pick)	LLM4Jobs: Achananuparp & Lim 2024 (SOC), Bach et al. 2025 (ISCO), Sturgis 2025 (SOC)	job title, optional description	35%-66% Top1(gpt03) = 0.74, Top1 43%-58%
Full fine tuning	LABOR-LLM (Athey et al. 2024); Safikhani et al. 2023; Kim et al. 2024	Survey answers (career histories, KWCS job+task sentences, DZHW job-title + task).	gpt3 - 30% BERT - up to 84%

We think we're swapping in a better model. We're actually making a dozen measurement decisions

Where the measurement decisions sit



Our study: 35 retrieval + 84 ranking experiments

A Separated Retrieval (embeddings) and Ranking (LLM); zero-shot prompting

B **Compared 5 different embedding models**
GPT:text- embedding-3- large
multilingual- MiniLM-L12- v2
Alibaba- NLP/gte- multilingual- base
IBM:granite- embedding- 278m- multilingual
Gemini:text- embedding- 004

C **Compared 7 different occupational component**s: Job descriptions, Tasks, Examples, and its combinations

D **Created 2 LLM personas** (occupational coding expert and respondent)

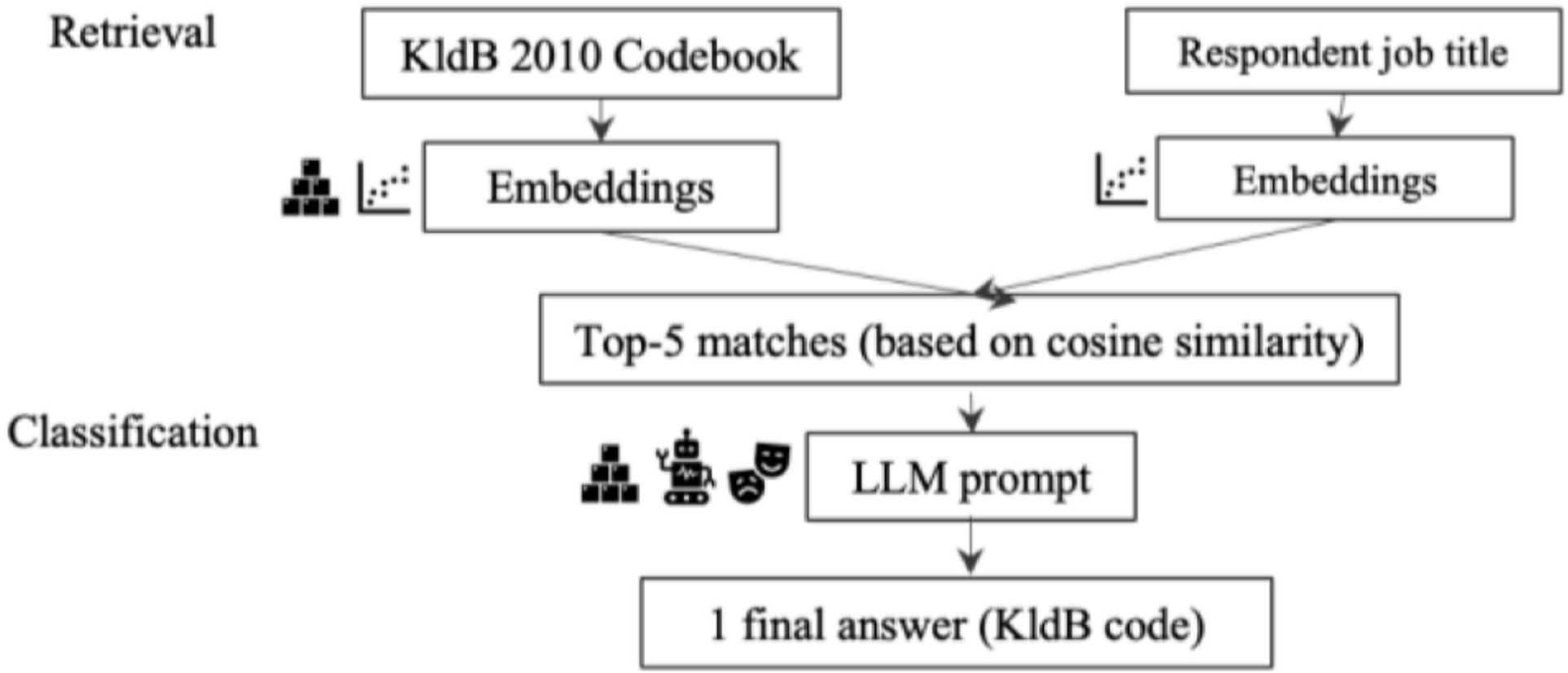
E **Compared 6 different LLMs** (GPT-4o-mini; Llama-3.1-70b-instruct; DeepSeek-R1; Mistral-small; Gemini-1.5-flas; Qwen2.5-vl-72b- instruct)

F **Compared 7 different occupational component**s: Job descriptions, Tasks, Examples, and its combinations

G Used population survey

Retrieval phase

- ✔ **How do different occupational information component**s (i.e. descriptions, tasks, examples, and their combinations) **and embedding models influence retrieval accuracy** when matching respondents' job titles to KldB 2010 occupational categories?


Retrieval
KldB 2010 Codebook
Respondent job title
Embeddings
Embeddings
Top-5 matches (based on cosine similarity)
Classification
LLM prompt
1 final answer (KldB code)

What text stands in for a category changes retrieval by 8–23 pp.

Compo-nent	GPT	Mini-LM	Ali-baba	IBM	Gem.
D	**62**	34	49	48	47
T	63	**30**	**39**	**32**	**39**
E	63	31	52	52	47
D+T	65	38	51	49	51
D+E	67	**40**	**57**	**55**	**58**
T+E	**69**	**30**	53	49	54
D+T+E	68	39	56	55	57

- 16% of answers were hard to retrieve consistently (e.g. geriatric nurse, clerk, engineer).
- 7% were always retrieved correctly across all setups (e.g. bank clerk, lawyer, journalist).

- **Combining components (e.g. Descriptions + Examples) boosts accuracy vs. single components.**
- GPT embeddings consistently outperform other models
- Respondent answer variability matters.

Classification phase

- ✔ How do **prompt structure** (occupational coding expert vs. survey respondent roles), **occupational information components**, and **choice of LLM model affect classification accuracy** in selecting the most appropriate occupational category among the Top 5 retrieved options?

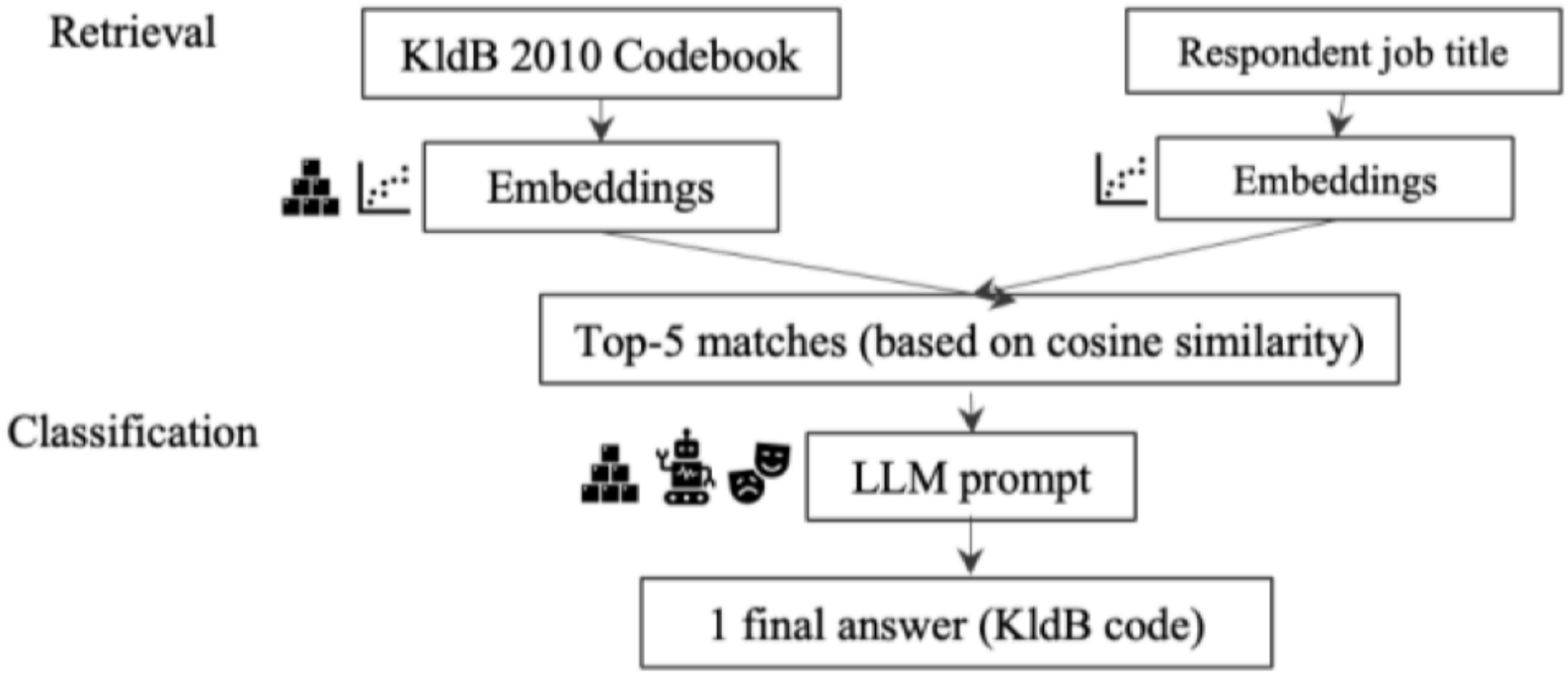

Retrieval
KldB 2010 Codebook
Respondent job title
Embeddings
Embeddings
Top-5 matches (based on cosine similarity)
Classification
LLM prompt
1 final answer (KldB code)

Table 1: Prompt structure

LLM as an Expert	**LLM as a Respondent**
You act as an experienced expert in occupational classification according to the German KldB 2010 system. Your in-depth expertise enables you to accurately assign job titles to the appropriate KldB categories. You will receive a job title along with a list of KldB categories – your task is to select the category that best describes the job title. Job title: [Job title] Please consider the following categories: • 1. [Description 1] • 2. [Description 2] • 3. [Description 3] • 4. [Description 4] • 5. [Description 5] • 6. None of the above categories Respond only with one of the following statements (please do not add any further explanations or text), followed by the corresponding number: - My final answer is: [Number] Example: "My final answer is: 3"	Imagine you are participating in a survey about your professional activity. You are [gender] and [age] years old. Your job title is: [job title]. Please read the following descriptions carefully and select the one that best reflects your job title. If multiple descriptions apply, choose the one that best describes your primary activity. If none of the descriptions fit well, select Option 6 ("None of the above categories"). • 1. [Description 1] • 2. [Description 2] • 3. [Description 3] • 4. [Description 4] • 5. [Description 5] • 6. None of the above categories Respond only with one of the following statements (please do not add any additional explanations or text), followed by the corresponding number: - My final answer is: [Number] Example: "My final answer is: 3"

Same construct, opposite optimum.

Role	Component	DeepSeek	Mistral	Gemini	GPT	Llama	Qwen
E	D	54	42	56	55	51	50
E	T	**49**	**36**	**47**	**46**	**44**	**43**
E	E	**64**	**50**	**64**	**64**	56	56
E	D+T	54	38	56	53	49	51
E	D+E	58	**50**	**64**	60	**58**	**57**
E	T+E	58	44	58	52	54	56
E	D+T+E	59	42	60	59	55	54
R	D	54	41	55	57	48	53
R	T	**50**	**30**	**44**	**46**	**42**	**44**
R	E	**64**	**50**	**65**	**63**	54	**61**
R	D+T	52	33	52	55	45	48
R	D+E	58	48	63	62	**56**	58
R	T+E	58	39	59	59	55	56
R	D+T+E	60	38	61	60	**56**	56

- Occupational component matters, **Examples** show strong performance
- All model underperform if **Tasks** or **Descriptions + Tasks** are used in the prompt

Expert coder' is a choice, not a default

- Respondent role returns higher accuracy scores
- Respondent prompt has a small average benefit, but for several models it make performance worse

- Examples or examples + descriptions are most beneficial occupational components to use in the prompt regardless the model
- GPT seems most robust, but DeepSeek and Gemini perform similarly well
- Respondent answer variability matters

Abstention: Identical instructions, incomparable behaviour

- Gold code absent in 33% of cases.
  With the same "None of the above" wording:
  GPT abstains correctly 45%,
  DeepSeek 44%,
  Llama 21%,
  Qwen 18%,
  Mistral 5%.

- **calibrate per model before deployment**

Conclusion

The ceiling: ICC = 0.76

The same model performs well or poorly depending on how categories are operationalised and how the prompt is framed

16% of jobs are always missed
Models can’t recover poor or ambiguous input. Better survey data is essential.

Smaller models need more targeted calibration with weights

Recommendation

choose representation per stage — compound for retrieval, examples for classification

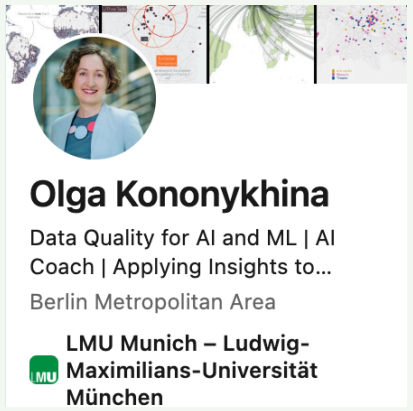
Test prompt role per model

calibrate abstention per model + deployment

invest at collection — the ICC ceiling won't yield to pipeline tuning

Pick a model, and you've made your measurement decisions whether you meant to or not

Thoughts? Questions?



You can always find me on Linkedin or
write to me: olga@bettermeasured.world