

LATENT INTELLIGENCE



Alan Berg

University of Amsterdam

a.m.berg@uva.nl

Stefan T. Mol, Tanja Hentschel

Amsterdam Business School, University of Amsterdam

Francesca Manzi

London School of Economics and Political Science

ALAN BERG

BACKGROUND

PhD Learning Analytics

25 Years Central IT Services University of Amsterdam
Data and Analytics expert.

Currently

AI4VET4AI – AI literacy in VET Education

Npuls – NL Digital Transformation for Education

Learning Analytics Team: Best and worst Practice

RESEARCH QUESTIONS

ANDROGENIZATION OF LEADERSHIP

IN THE JOB MARKET

Hypothesis 1: Regardless of time, characterizations of leadership are more agentic than communal. Specifically, Agentic words are used more frequently than communal words in leadership job ads.

Hypothesis 2: Communal characterizations of leadership have increased over time. Specifically, the use of communal words in leadership job ads has increased over time.

Hypothesis 3: The increase in communal words is strongest in female dominated industries, followed by gender balanced industries, and male dominated industries.

METHOD

UK dataset 2014:2020
many Millions of
vacancies from
Burningglass

Binary classification – Leadership/Not Leadership

Measuring change over time via **dictionaries**

Generic pipeline

- Reuse for other populations.
- Extendable (entity and relationship extraction)
- Allows Domain experts to write their selection criteria in plain text.

Scales / Sustainable /




Explainable / Extendable

Runs on laptop

WORKFLOW

- 1) Design Research questions and preregister questions and method
 - 2) Develop Benchmark (small sample set)
 - 3) Design and improve prompts based on benchmark
 - 4) LLM (8 billion) creates dataset (20,000 – 10,000 per laptop day)
 - 5) ML models compared (Explainability vs Performance tradeoff).
 - 6) ML Model characterizes whole Population
 - (10,000 times faster than LLM on my laptop)
 - 7) Dictionary counts
 - 8) Statistical analysis
 - 9) Publish – Peer review
-

LOOSE COUPLING

- 1) **Design Research questions and preregister questions and method**
 - 2) **Develop Benchmark (small sample set)**
 - 3) **Design and improve prompts based on benchmark**
 - 4) LLM creates dataset (20,000 – 10,000 per laptop day)  **Explainability**
 - 5) ML models compared.
 - 6) ML Model characterizes whole Population  **Explainability, Scalability**
 - (10000 times faster than LLM)
 - 7) Dictionary counts  **Explainability, Scalability**
 - 8) **Statistical analysis**
 - 9) **Publish – Peer review**
-

FEATURE IMPORTANCE

Agency words impacting characterisation

term	Details	
	estimate	Rank
leadership	326.76923	3
lead	301.01391	4
managerial	160.96201	9
leader	160.65894	10
manager	141.74829	17
leading	125.25585	28
acceptable	74.35334	75
applicable	69.83255	96
controlling	54.56566	162
productive	52.04207	182
logical	51.28350	188
managers	50.35548	194
achieve	49.21029	200
effective	48.63825	203

Communion words impacting characterisation

term	Details	
	estimate	Rank
contacted	73.44009	79
helped	54.87256	159
honesty	53.00846	175
affect	52.98551	176
opened	51.51535	185
helpful	50.54000	192
tactical	46.92002	222
sense	-40.40348	4703
loves	-40.76372	4708
sharepoint	-43.06292	4743
communications	-53.18858	4854
contacting	-53.65742	4855

Top impactful words that are not Agency or Communion

term	Details	
	estimate	Rank
supervise	343.4215	1
supervising	341.0540	2
017	253.1392	5
atlanta	213.0602	6
supervisory	175.0583	7
mentoring	161.3096	8
mentor	157.1176	11
personalise	149.8899	12
supervisor	146.7656	13
motivate	145.6976	14
manage	144.7025	15
sees	143.5900	16
adverts	140.6161	18
jobs.ie	139.3467	19
overseeing	137.2621	20
toolbox	135.7567	21
managing	134.6575	22
7,000	131.9762	23
subcontractors	128.5285	24
supervision	127.3571	25

Note

Words later removed from dictionaries that related to job titles

EXPLAINABILITY

- Coding book
- Prompt
- Benchmarked data
- Ask LLM to explain

The job description emphasizes staff management, training, and workflow maintenance, which suggests a leadership role. The responsibilities also include decision-making, such as dealing with QOF alerts and ensuring the reception area is maintained to high standards. This indicates a level of autonomy and authority typically associated with leadership positions.

- Feature Importance from ML model
 - Reuse of peer reviewed Dictionary
 - Open-Source Model that is also described in research papers
-

OPERATIONALIZATION

1) Design Research questions and preregister questions and method

2) Develop Benchmark (small sample set)

3) Design and improve prompts based on benchmark

4) LLM creates dataset (20,000 – 10,000 per laptop day)

5) ML models compared.

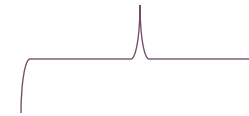
6) ML Model characterizes whole Population
○ (10000 times faster than LLM)

7) Dictionary counts

8) Statistical analysis

9) Publish – Peer review

PROMPTS



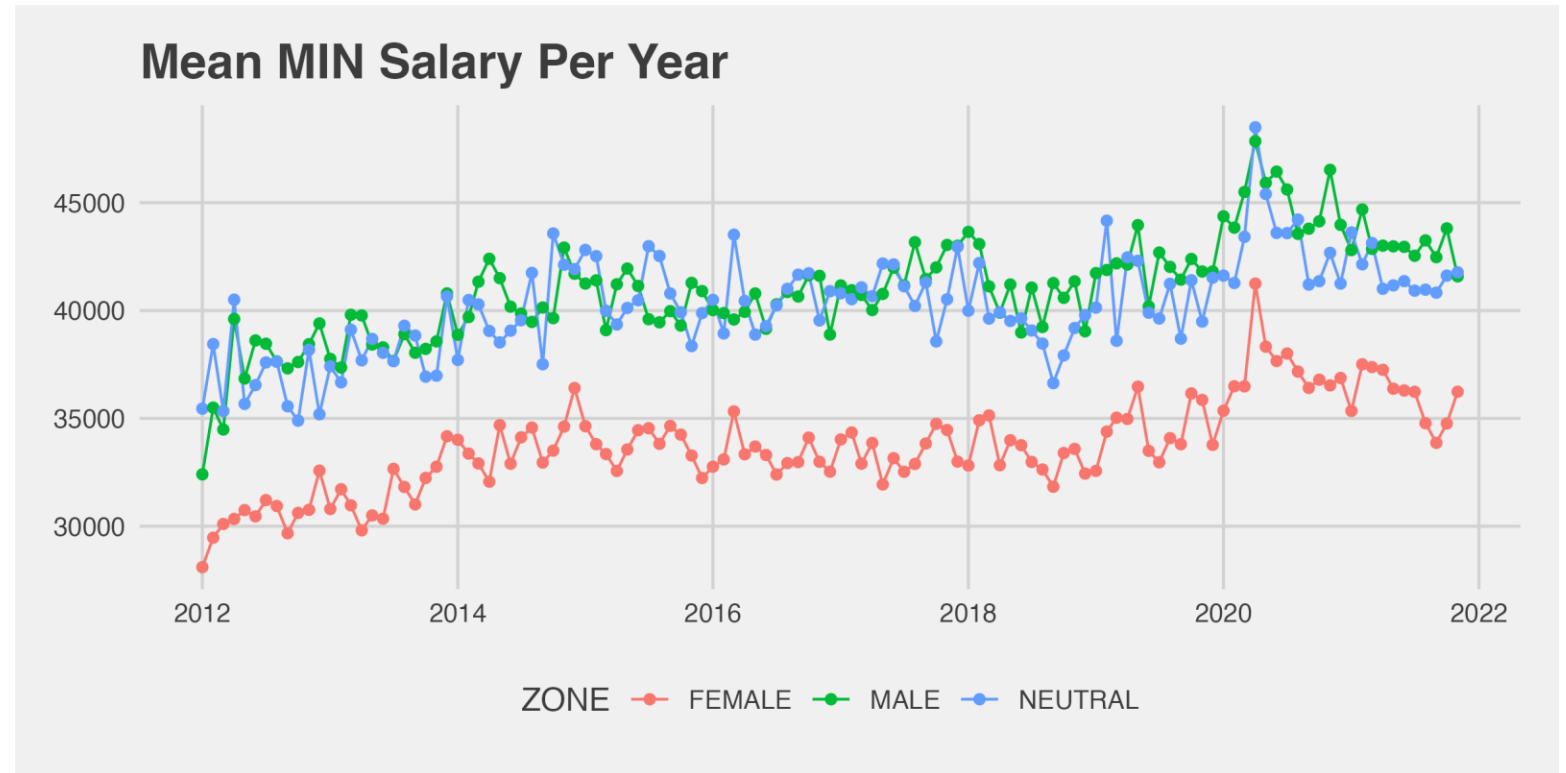
ENTITY AND RELATIONSHIP
EXTRACTION AND RECONCILIATION

- EARLY WARNING TAXONOMY CHANGES
- DATASET for DISTILLATION TO SMALL LM

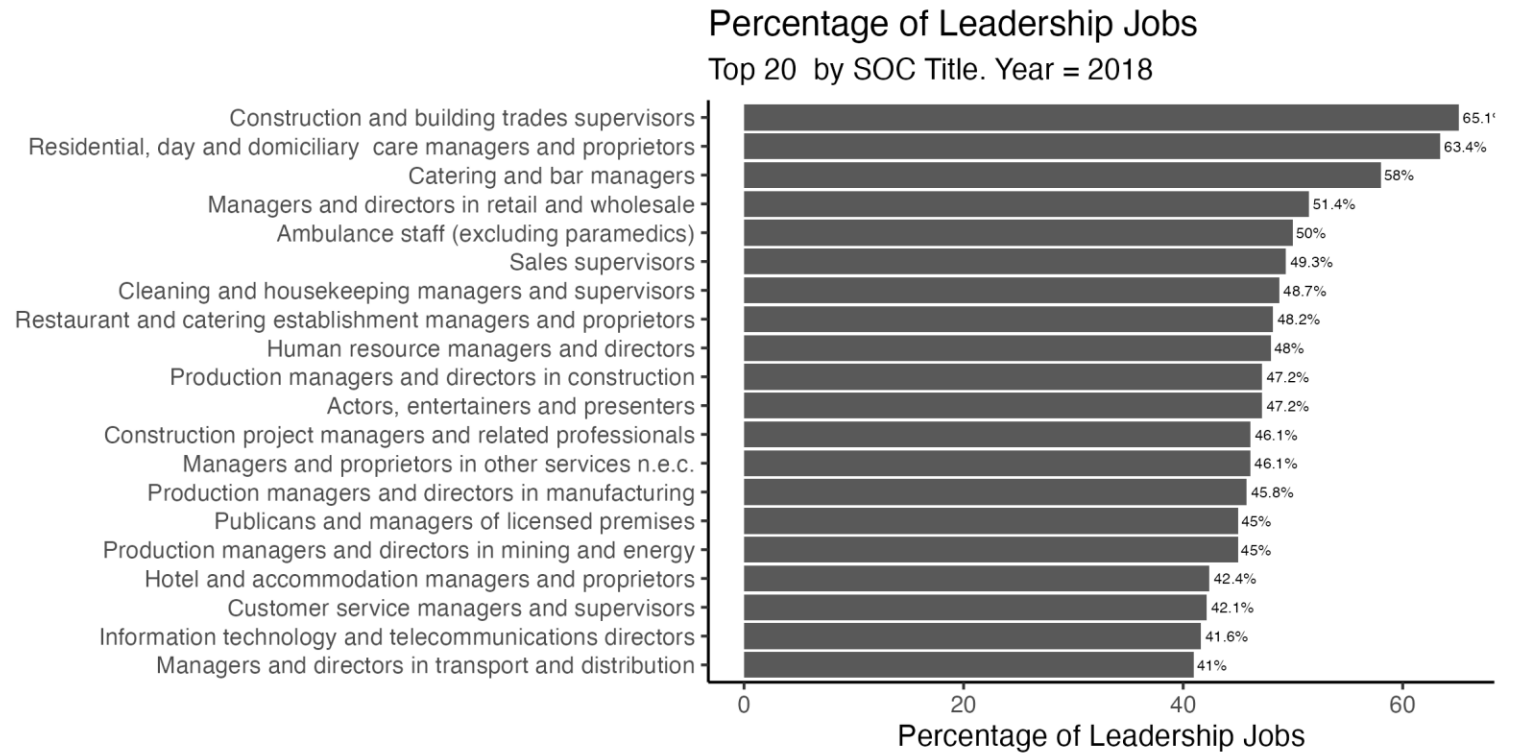


LLM / ML CYCLE

CALIBRATE WITH KNOWN FACTS



CALIBRATE WITH KNOWN FACTS



EXTRA (PROBABLY NOT PRESENTATION)

SCALABLE RESEARCH

My unpaid opinion...

Slides 13 onwards not presented.

LOOSE COUPLING OF RESPONSIBILITIES

- 1) **Design Research questions and preregister questions and method**
 - 2) **Develop Benchmark (small sample set)**
 - 3) **Design and improve prompts based on benchmark**
 - 4) LLM creates dataset (20,000 – 10,000 per laptop day)
 - 5) ML models compared.
 - 6) ML Model characterizes whole Population
 - (10000 times faster than LLM)
 - 7) Dictionary counts
 - 8) **Statistical analysis**
 - 9) **Publish – Peer review**
-

CONSTRAINTS

Working with **domain experts** that can describe leadership

Potentially Multilingual vacancy **data**

Take advantage of **using LLM's**

Communicate in plain text

Multilingual

Improving rapidly (free development)

Runs on laptop

Easy to translate domain specific questions into characterization

Cheaper than human rankers

I am not rich. My research is an unpaid mandate. However, I own a laptop and can program in R and **open-source** models

Do not want to contribute to **climate change**

Need to **explain outcomes**

THOUGHT EXPERIMENT

LAZY COUPLING OF EXPERTISE

Provide:

- Code to run locally (R or Python notebook)
- Installation instruction
- Prompt design instructions
- Github location to share results

Ask for:

- Research Questions
- Entities to extract
- Benchmark of provided dataset sample
- Relationships with hierarchy to extract
- Prompts

Return:

- Aggregated results

Domain experts' play to their strengths.
Effort is diminished
Easy access to the aggregated dataset
Uniform method that aids in
comparison
Visibility between experts

Entity and relationship extraction for
taxonomy maintenance

You are responsible for :

- optimizations based on EC values.
- LLM opts
- Promoting to shared statistics
- Early warning of relationship changes in entities such as new skills.
- Administration and improvement

Provide:

- Code to run locally (R or Python notebook)
- Installation instruction
- Prompt design instructions
- Github location to share results

Ask for:

- Research Questions
- Entities to extract
- Benchmark of provided dataset sample
- Relationships with hierarchy to extract
- Prompts

Return:

- Aggregated results

REFLECTION

1. Are you doing this already
Is there a method for taxonomy generation coupled with domain experts that is scalable.
2. Is scalable even needed with sampling
3. Past skills, which entities and relationships do you want to extract?
4. Are we serious about sustainability. Perhaps a CO2 estimator is needed?
5. Which other EC defined values can be supported.
6. What is the minimum amount of explainability needed.
7. Do we need to debias national LLM's such as DeepSeek R1
8. Is it worth fine tuning, distilling models to cheaper smaller more sustainable versions
9. How do we keep up with state of the art
10. Do we need to decide to embrace Open-Source vs Close Sourced models
11. How do we generate shareable benchmarks and annotated (human,AI) datasets?
12. How do we methodically test concept validity