LLM-Based Occupational Risk Measurement

*CPS + O*NET*

Alec Isaacman
ECON 542

Motivation

Why do we care about occupational risk?

- Occupational risk matters for wages, job security, and policy
- Standard datasets do not directly measure risk exposure
- I use an LLM to construct interpretable risk measures from text

Occupation Text → LLM → Risk Labels → Worker-Level Data

Data Sources

O*NET Occupation Data

- Titles + descriptions (1,016 SOC occupations)

IPUMS CPS Microdata

- ~834k individual workers

Census OCC → SOC Crosswalk (.csv)

- Required to merge CPS to O*NET

Occupations: 1016

	soc	title	description
0	11-1011.00	Chief Executives	Determine and formulate policies and provide o...
1	11-1011.03	Chief Sustainability Officers	Communicate and coordinate with management, sh...
2	11-1021.00	General and Operations Managers	Plan, direct, or coordinate the operations of ...
3	11-1031.00	Legislators	Develop, introduce, or enact laws and statutes...
4	11-2011.00	Advertising and Promotions Managers	Plan, direct, or coordinate advertising polici...

```
print("Observations:", cps.shape[0])
print("Variables:", cps.shape[1])
```

```
Observations: 406632
Variables: 39
```

Measuring Risk

What risks are measured?

Three dimensions:

- **Physical risk**
- **Financial liability risk**
- **Cyclical job security risk**

Each classified as:

- Low
- Medium
- High

Risk Classification Prompt

```
SYSTEM_PROMPT = """
You are an economist labeling occupational risk.
You must return ONLY a valid JSON object.
No markdown. No commentary. No explanation.
If unsure, still choose Low, Medium, or High.
"""


USER_PROMPT_TEMPLATE = """
Given the occupation below, classify risk levels.

Return JSON with exactly these keys:
- physical_risk
- financial_liability_risk
- cyclical_job_security_risk

Each value must be one of: Low, Medium, High.

{llm_text}
"""
```

Example LLM Output

LLM model applies Low / Medium / High risk to occupation title + description

risk_test.head()

	soc	title	physical_risk	financial_liability_risk	cyclical_job_security_risk
0	47-4041.00	Hazardous Materials Removal Workers	High	Medium	Medium
1	31-9093.00	Medical Equipment Preparers	Medium	Low	Low
2	19-1022.00	Microbiologists	Low	Medium	Medium
3	53-1043.00	First-Line Supervisors of Material-Moving Mach...	Medium	Medium	Medium
4	51-9061.00	Inspectors, Testers, Sorters, Samplers, and We...	Medium	Low	Medium
5	33-2022.00	Forest Fire Inspectors and Prevention Specialists	High	Medium	Medium
6	49-9061.00	Camera and Photographic Equipment Repairers	Medium	Low	Medium
7	17-3027.00	Mechanical Engineering Technologists and Techn...	Medium	Medium	Medium
8	19-1029.03	Geneticists	Low	Medium	Medium
9	39-3012.00	Gambling and Sports Book Writers and Runners	Low	High	Medium
10	39-3093.00	Locker Room, Coatroom, and Dressing Room Atten...	Low	Low	Medium
11	29-1127.00	Speech-Language Pathologists	Low	Medium	Low
12	33-9031.00	Gambling Surveillance Officers and Gambling In...	Medium	High	Medium
13	39-4012.00	Crematory Operators	Medium	Low	Low

Full O*NET Risk Distribution

1,016 occupations labeled

```
risk_onet["physical_risk"].value_counts()
```

physical_risk	count
Low	447
Medium	347
High	222

The Hard Part: OCC → SOC Crosswalk

- **CPS uses OCC codes; O*NET uses SOC**
- **Extensive cleaning required**
- **Used official Census crosswalk (2018+) (.csv)**
- **Cleaned, deduplicated, and merged**

```
CPS rows: 406632  (was 834419 )
SOC missing share: 0.21567412303999686
```

	occ	soc
0	6200	471011.0
1	4720	412010.0
2	7700	511011.0
3	205	119013.0
4	2145	232011.0

```
Clean rows: 472
Unique OCC: 472
Unique SOC: 472
```

	occ	soc	title
0	0	0	Not Applicable (Under 16 years or not in the l...
1	20	111021	General and Operations Managers
2	40	112011	Advertising and Promotions Managers
3	51	112021	Marketing Managers
4	52	112022	Sales Managers
5	60	112030	Public Relations and Fundraising Managers
6	101	113012	Administrative Services Managers
7	102	113013	Facilities Managers
8	110	113021	Computer and Information Systems Managers
9	120	113031	Financial Managers

Merging Risk into CPS

	occ	soc	soc6	physical_risk	financial_liability_risk	cyclical_job_security_risk
0	6200	471011	471011	High	Medium	High
1	4720	412010	412010	NaN	NaN	NaN
2	7700	511011	511011	Medium	Medium	Medium
3	205	119013	119013	High	Medium	High
4	2145	232011	232011	Low	Medium	Medium
5	8320	516031	516031	Medium	Low	High
6	205	119013	119013	High	Medium	High
7	205	119013	119013	High	Medium	High
8	4510	395012	395012	Medium	Low	Medium
9	205	119013	119013	High	Medium	High

Final Risk Measures in CPS

- **Risk labels converted to numeric scale (0–2)**

- **Enables descriptive statistics and regression use**

	physical_risk_num	financial_liability_risk_num	cyclical_job_security_risk_num
count	240408.000000	240408.000000	240408.000000
mean	0.788235	0.712618	1.158148
std	0.801959	0.606631	0.577341
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	1.000000
50%	1.000000	1.000000	1.000000
75%	1.000000	1.000000	2.000000
max	2.000000	2.000000	2.000000

Notable Findings

High physical risk jobs ≈ 24%

High cyclical risk jobs ≈ 26%

```
analysis_df[["phys_high", "cycle_high"]].mean()
```

	0
phys_high	0.238108
cycle_high	0.258240

dtype: float64

Limitations

- **Risk measured at occupation level, not individual level**
- **Partial SOC coverage**
- **LLM judgments depend on text quality**

Takeaways

- **LLMs can generate interpretable economic measurements**
- **Pipeline integrates text → risk → microdata**
- **Framework is extensible to wages, safety, policy analysis**

Moving forward:

Can apply this framework to the construction industry to enhance my wage premium analysis.