**Questions for customers 09/10**

Which skills needed?

* Hard\_skills
* Soft\_skills?
* General skill-terms? (Data Strategy & Roadmap Development, Project Management, End-user Training, ...)
* How to handle new skills (llm only risky, “new\_pool” suggest documentation of new skill based on threshold #occurrences)

Scope of job applications?

* Consulting
* IT
* HR?
* …

Prepared to pay for llm-API for historical data processing?

* Cost estimation = 3eur / 6500 records (openAI batch API)
* Local llm will take approximately 23 full days to process 6500 records
* Local llm will take 2-3hrs / 20 records (faster with quantized models?)

Extra columns: extracted\_skills (language requirements?), skill\_proficiency\_levels, skills\_group, skills\_subgroup, language, week\_number

Potential columns?: llm\_model\_used, llm\_version\_date, num\_skills\_detected, new\_skill(bool),

PLAN:

Design vector embedding (+ synonyms) for general skills

Enrich historical data (TUNING LLM for openAI batch API !)

* Extract (3) extracted\_skills, skill\_proficiency\_levels, language from description, skill\_synonyms.json
* Extract (2) skill\_group, skill\_subgroup from skill\_to\_info.json
* Extract (1) week\_number from date\_posted

Orchestrate enrichment for future data (local llm?)

* Dagster (customer or us?)

Enrich skill taxonomy with historical data (requirements for acceptance?)

Design futureproof “new skills handling system” - proposal

* New skills pool with #occurrences (synonyms will occur)
* Set occurrence threshold
* Suggest new skill documentation (add to skill\_to\_info.json, go through new skills pool for synonym detection and add to skill\_synonyms.json)

Orchestrate “new skills handling system”

* Dagster