LLM-Assisted Data Extraction

Jared Edgerton

Why Use LLMs for Data Extraction

Many social-science data sources are:

- Messy
- Unstructured
- Text-heavy
- Expensive to code by hand

LLMs offer scalable *assistance*, not automation.

Human-in-the-Loop Philosophy

In this course, LLMs are used to:

- Propose structured outputs
- Accelerate labeling
- Surface uncertainty

Humans remain responsible for validation.

Structured Extraction

Structured extraction means:

- Defining a schema in advance
- Forcing outputs into fields
- Rejecting free-form text

Schemas discipline model behavior.

Schema Example

```
schema = {
    'actor': 'string',
    'action': 'string',
    'target': 'string',
    'date': 'YYYY-MM-DD',
    'confidence': 'float'
}
```

Explicit schemas reduce ambiguity.

Prompt Design for Extraction

Effective prompts:

- Describe the task narrowly
- Specify output format
- Include examples when possible

Prompting is part of the method.

Prompt Example

```
prompt = '
Extract the following fields from the text below.
Return JSON matching this schema:
{schema}

Text:
{text}
'
```

Batch Processing

LLMs are typically applied:

- In batches
- With rate limits
- With cost constraints

Pipelines must manage scale.

Batch Processing Pattern

```python
results = []
for doc in documents:
    response = call_llm(prompt_template, doc)
    results.append(response)
```

Batch size affects cost and reliability.

Uncertainty and Confidence

LLMs can:

- Provide confidence scores
- Flag ambiguous cases
- Abstain when uncertain

Uncertainty should be captured explicitly.

Uncertainty Example

```
if response['confidence'] < 0.6:
    flag_for_review(response)
```

Low-confidence cases are routed to humans.

Human Validation

Human review is used to:

- Spot-check outputs
- Correct systematic errors
- Refine prompts and schemas

Validation is iterative.

Spot Audits

Auditing strategies include:

- Random sampling
- Stratified checks
- Edge-case review

Audits reveal failure modes.

Failure Modes

Common failures include:

- Hallucinated fields
- Overconfident errors
- Inconsistent formatting
- Sensitivity to phrasing

Failures must be documented.

Evaluation Metrics

Extraction quality can be assessed with:

- Precision / recall
- Field-level accuracy
- Agreement with human labels

Evaluation depends on task goals.

LLMs Are Not Neutral

LLM outputs reflect:

- Training data biases
- Prompt framing
- Model defaults

Human oversight is essential.

Documentation Requirements

Every LLM pipeline should record:

- Model and version
- Prompt text
- Schema definition
- Validation procedure
- Known limitations

Transparency enables reuse.

What We Emphasize in Practice

- Use LLMs as assistants
- Enforce structure aggressively
- Validate with humans
- Document failure modes

Discussion

- Where do LLMs help most?
- When do they fail quietly?
- How much validation is enough?