Learning to Make Documents Barrier-free and Accessible

Benchmarking Question and Testing Framework (HOFFMAN*)

***JKU AI Masters Practical Work**

Jack Heseltine **for IT:U NLP Group - Pres. 1 | 15 mins with Q&A**

JOHANNES KEPLER UNIVERSITY LINZ

Introduction

Coming from a Masters Thesis project at JKU, the ultimate goal is the annotate PDFs for people using screenreaders = multimodal, document-centric task, **limited training-data**

Of interest because of this last point: **few-shot paradigm**, generally speaking, and also the emergent phenomena idea in LLM performance - all background to this project

Present **Practical Work: Setting up a testing framework** to try various hyper-parameters. Current progress available on Overleaf

Outlook: **Pres. 2**, on Linguistic Competence - I will get to this question, but it seems I am in the more usual frame of a hard, knowledge-based, verifiable (PDF-checker tools) domain, albeit with soft, linguistic, hard-to-check aspects, e.g. quality of an image summary (on a textual level)

JOHANNES KEPLER UNIVERSITY LINZ

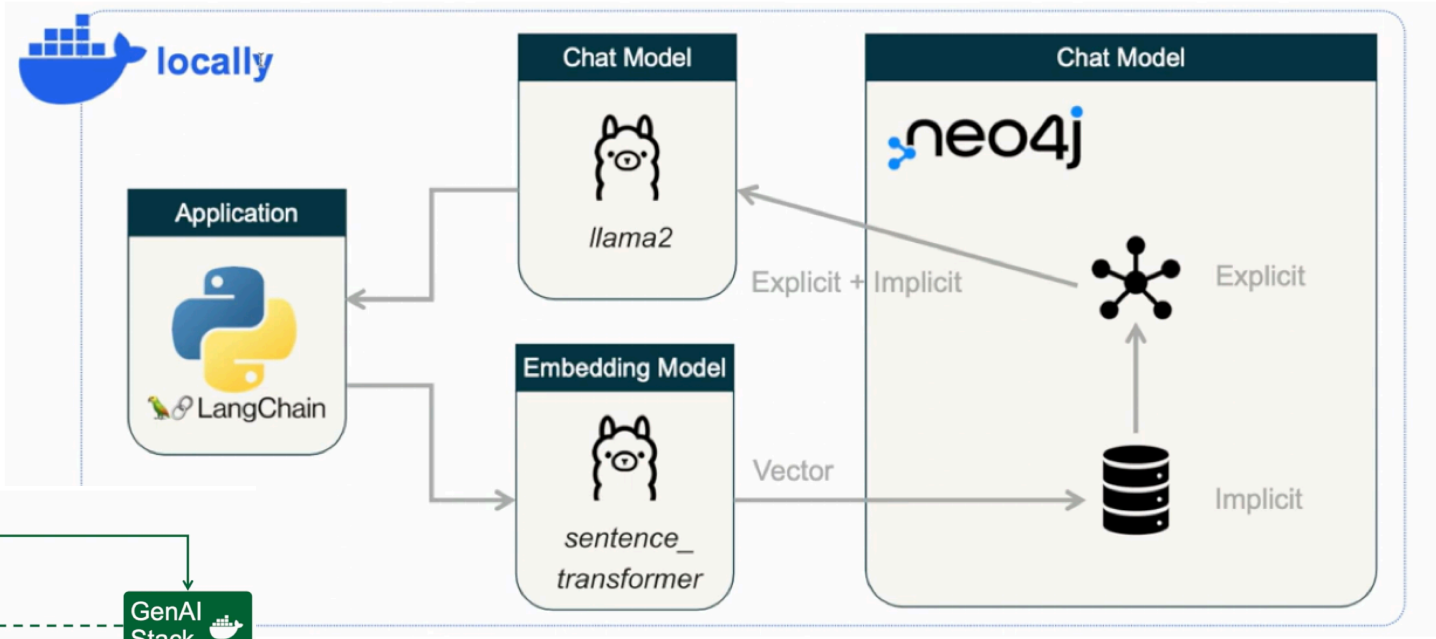
JKU LINZ INSTITUTE OF TECHNOLOGY

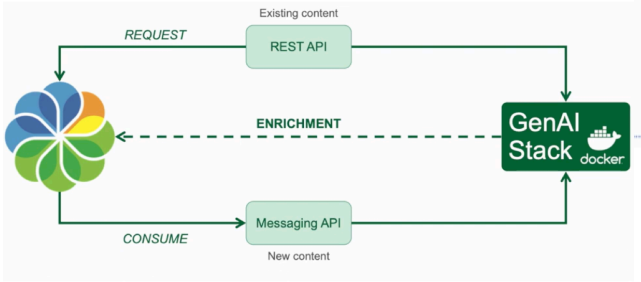
FAW.

HOFFMAN: Let's Start with the Testing Architecture

DEMO/Overview:

Or circle back
to Demo as needed





JOHANNES KEPLER UNIVERSITY LINZ

LINZ INSTITUTE OF TECHNOLOGY

UpdateSocial an AI research project @ccessibility buddies

FAW.

Theory Background: Few-Shot/Unsupervised Learners

I have a whole **seminar presentation (and paper) on this topic**

Of interest in 2025: **critiques** of the emergent abilities* idea, e.g. Lu et al., 2023 (Are Emergent Abilities in LLMs just In-Context Leaning?)

- in-context learning
- model memory
- linguistic knowledge

*Introduced as abilities in LLMs that are absent in smaller models (same data assumption) - from physics

Connects immediately with Mahowald et al., 2023 (Dissociating Language and Thought in LLMs) - i.e. formal vs functional linguistic abilities as a further confounding factor in the above view

Bottomline? To not overestimate LLM-capabilities, most likely. (Practical bottomline will follow as well.)

Document-Focus: Transformation D —> D`

Firmly in the formal language capability domain to begin with, as per the previous slide, looking for an **appropriate D'**. Question is how to evaluate: in this I find the question does lead into functional language considerations, necessarily,

Three approaches: **first**, document checking (PDF-checker) - this needs to be fulfilled so the **document is readable** and usable for its intended purpose

Second, a **gradient of further quality aspects** improving screenreader user experience gradually, like fully nesting structure of hierarchy annotations and the like

Third, I would like to add a **linguistic layer** capturing the open-ended aspects like summary (textual) quality and alt-texts generally

I would like to summarize these three factors into a **score card or traffic light**, speaking to PDF failure as well as A-D/ 1-4 grade quality, **aligning this score with actual user feedback**. For this part I am in touch with the JKU Institute of Integrated Studies. It might very well be the case that an A-score would require linguistic competence performance.

JKU JOHANNES KEPLER UNIVERSITY LINZ

JKU Institute Integrated Studies

Document-Focus: Transformation D —> D`

1. Minimal PDF Without Accessibility Annotations

plaintext Copy code

```
%PDF-1.7
1 0 obj
<<
  /Type /Catalog
  /Pages 2 0 R
>>
endobj
2 0 obj
<<
  /Type /Pages
  /Kids [3 0 R]
  /Count 1
>>
endobj
3 0 obj
<<
  /Type /Page
  /Parent 2 0 R
  /MediaBox [0 0 612 792]
  /Contents 4 0 R
>>
endobj
4 0 obj
<<
  /Length 44
>>
```

2. PDF with Accessibility Annotations (PDF/UA Compliant)

plaintext Copy code

```
%PDF-1.7
1 0 obj
<<
  /Type /Catalog
  /Pages 2 0 R
  /MarkInfo << /Marked true >>
  /StructTreeRoot 5 0 R
>>
endobj
2 0 obj
<<
  /Type /Pages
  /Kids [3 0 R]
  /Count 1
>>
endobj
3 0 obj
<<
  /Type /Page
```

A11y enhancements

- **Tags** /MarkInfo << /Marked true >>
- Structure Elements /S, /P, /H1, …
- Mapping Content to Tags
- Readable **Text**
- Logical **Reading Order**

Document-Focus: Checking (First level)

This is where I started and is the most straightforward: one core idea (implementation is currently in progress) is to build out a **classifier that is trained alongside a standard (non-neural-net) PDF-checker** for native checking capabilities in the framework, not relying on third-party tool. (This is submitted with the practical work component.)

PDF-checker used: **Adobe Developer PDF Accessibility Checker** (API, see below overview) - Idea: take this report and perform the previously manual work with LLM, potentially in an agentic/RAG framework and evaluate output

PDF Accessibility Checker

The Accessibility Checker API verifies if PDF files meet the machine-verifiable requirements of **PDF/UA and WCAG 2.0.** It generates a **report** summarizing the findings of the accessibility checks. Additional **human remediation** may be required to ensure the reading order of elements is correct and that alternative text tags properly convey the meaning of images. The report contains links to documentation that assists in manually fixing problems using Adobe Acrobat Pro.

Document Focus: RAG (and Agentic/Multiple Tries)

As shown on the architecture level I am already working in the LLM framework package LangChain and have document vector database support: the idea is to make use of **current trends in LLM tooling**, namely **RAG and agents** for, say, trying multiple approaches and incorporating PDF checker feedback, for instance.

(A practical bottomline of the theory introduced: using modern approaches and for this task, it does not matter if solving the task is emergent or not, actually, but nice to explore theoretically.)

I see this as the long-tail of the work once the central evaluation question has be tackled, so that I can work of **performance improvement in a verifiable way** and on this software level. If done inside a **framework that reliably tracks and integrates these various strands of development**, a qualitative outcome should be achieved, proving (or disproving) the concept with this set of tools. This is the **contribution of the Hoffman stack** and the goal with the thesis.

LLM Hard Metrics for Task-Agnostic Quality-Checking

PDF-Checking is task-specific: I am currently learning about the general way to check LLM transformation tasks, assuming available training data

I want to consider this in thinking about level two, training on PDF-checker/classification models as well as A-score samples.

Score something like **Quality Score = w1 * Text Accuracy + w2 * Tagging Precision + w3 * Reading Order Accuracy + w4 * Compliance**, where w4 would need to be heavily weighted

+ w5 * LLM Soft Metrics/Linguistic Performance? (Cf. **Pres. 2**)

LLM Hard Metrics for Task-Agnostic Quality-Checking

Text Accuracy

Metric Idea: Token-Level Accuracy, Levenshtein Distance or similar

+ Semantic Accuracy via BLEU (Bilingual Evaluation Understudy) or ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Tagging Precision, and Recall/F1-Score

Idea: Treat tagging as a classification problem …

True Positive (TP): Correctly predicted tags.

False Positive (FP): Predicted tags that do not exist in the ground truth.

False Negative (FN): Tags present in the ground truth but not predicted.

Reading Order Accuracy

Compare the logical reading order of the predicted text (determined by the structure tree) against the ground truth + use sequence alignment algorithms like the Needleman-Wunsch or Smith-Waterman algorithms.

A finding might also be that some or all of these are not suitable for evaluation of accessible document transformations

Masters Thesis Outlook (and Timeline)/Q&A

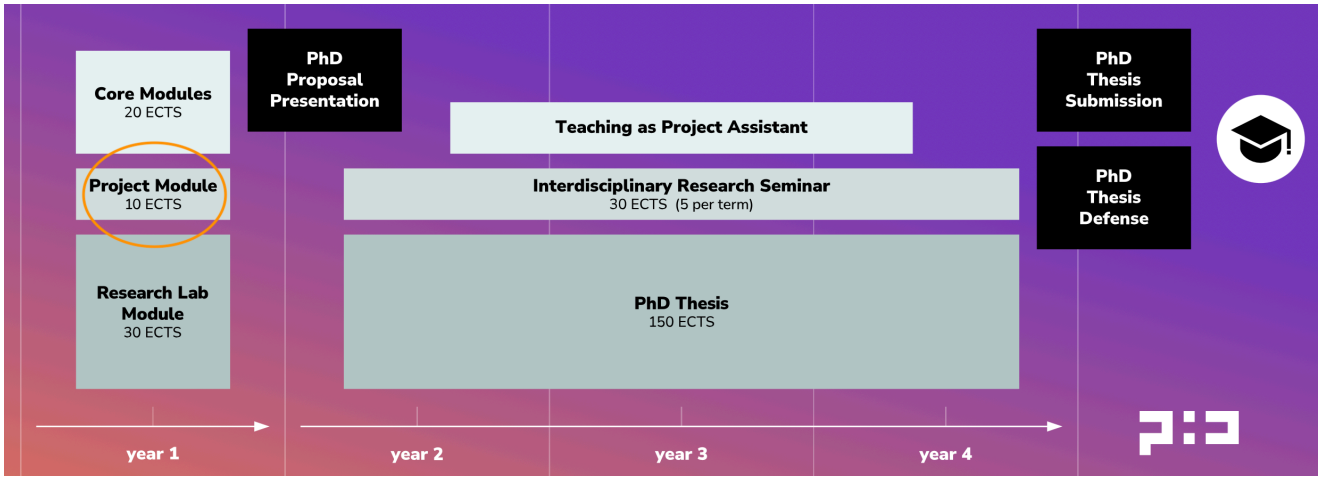
Now: Complete Practical Work and incorporate into thesis, abstract and structure already approved.

Testing focus will be evaluation in this framework, writing focus presenting the relevant background.

After **completing my thesis (FAW) in late spring 2025**, there are *likely still further topics and applications oriented questions I would like to explore (see research agenda discussion at the end of today's meeting and/ or blog note)* + ***ICCHP (International Conference on Computers Helping People with Special Needs) Young Researchers' Consortium 2026 paper goal*** *in collaboration with Institute of Integrated Studies at JKU*

(Final JKU AI exam June/July 2025.)

Q&A



JOHANNES KEPLER UNIVERSITY LINZ FAW

References

Full presentation and context: https://heseltime.github.io/rDai#it-u

Third-party references throughout slides