# Suitability of Large Language Models for Making PDF-Documents More Accessible and Barrier-free in Enterprise Content Management

Jack Heseltine

JOHANNES KEPLER UNIVERSITY LINZ

# Presenting today

- Master Thesis Work
  - Seminar: In-context Learning Papers
  - Practical Work: ECM Integration (Applied Solution)
  - **Thesis**: **LLM Fine-tuning and related experiments**
- Topic: LLMs to **Generate PDF Source Code** (Representation Format) with Annotations, Tags, … that make the file more readable for screenreader interpretation
- Legal relevance in 2025, generally an ECM topic/was considered in this (technical) context especially - some notes will follow

- Project assumptions subscribe to this formula [1]:

$$\boxed{\text{Accessibility} + \text{Usability} + \text{User Centered Design} = \text{Quality for All}}$$

  - [1, p.37]: Klaas Posselt and Dirk Frölich. 2019. Barrierefreie PDF-Dokumente erstellen. ISBN: 978-3-86490-487-5.

# Basics/Motivation

**Inclusion, access, barrier-freedom.** Barrier-freedom is mostly used in German language settings and means the “creation of a context that allows people the equal-rights, unhindered access to all areas of life” [1, p. 33] and is therefore crucial for real inclusion, the “independent, equal-rights participation of all people in social life” [1, p. 34], though it goes further than just the social sphere: access is the actual mechanism by which inclusion and barrier-freedom take place, in the present author’s definition. In the technological context, accessibility is mediated and extended by usability and user-centered design, introducing the fundamental aspect of quality, so we would also subscribe to the formula

$$\boxed{\text{Accessibility} + \text{Usability} + \text{User Centered Design} = \text{Quality for All}}$$

- [1, p. 37]: Klaas Posselt and Dirk Frölich. 2019. Barrierefreie PDF-Dokumente erstellen. ISBN: 978-3-86490-487-5.

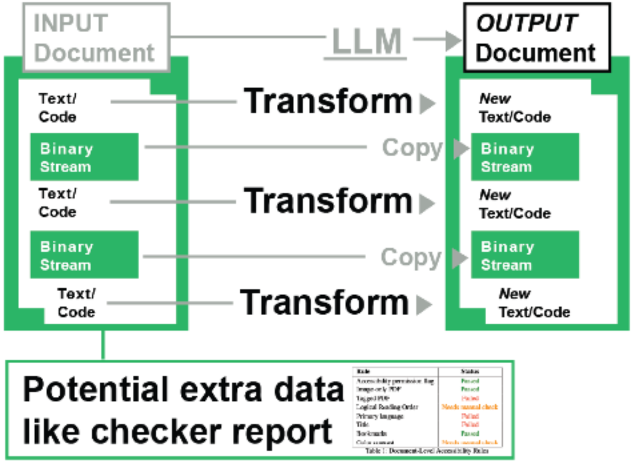
# Outline

- Core Challenge/Problem — Why is this an ML Topic?
- The Setting and Technical Situation
- Current Legal Context
- ECM Implementation (Part I - Not Focus)
- In-context Learning, Fine-tuning and Meta-Information Approaches (Part II - **Focus**)
- Disadvantages of the Chosen Approaches, (Current Work & Benchmark:) Final Experiments to Improve Scores
- **Results** for this work
- OOD Metrics
- NLP Measurements Used
- Conclusion and Outlook

# Core Challenge/Problem — Why is this an ML Topic?

- **Screen Readers: Demo 1**
- **Document Transformation**

… with varying objectives




## Accessibility Report

### Summary

The checker found problems which may prevent the document from being fully accessible.

- Needs manual check: 2
- Passed manually: 0
- Failed manually: 0
- Skipped: 0
- Passed: 11
- Failed: 19

### Detailed Report

**Document**

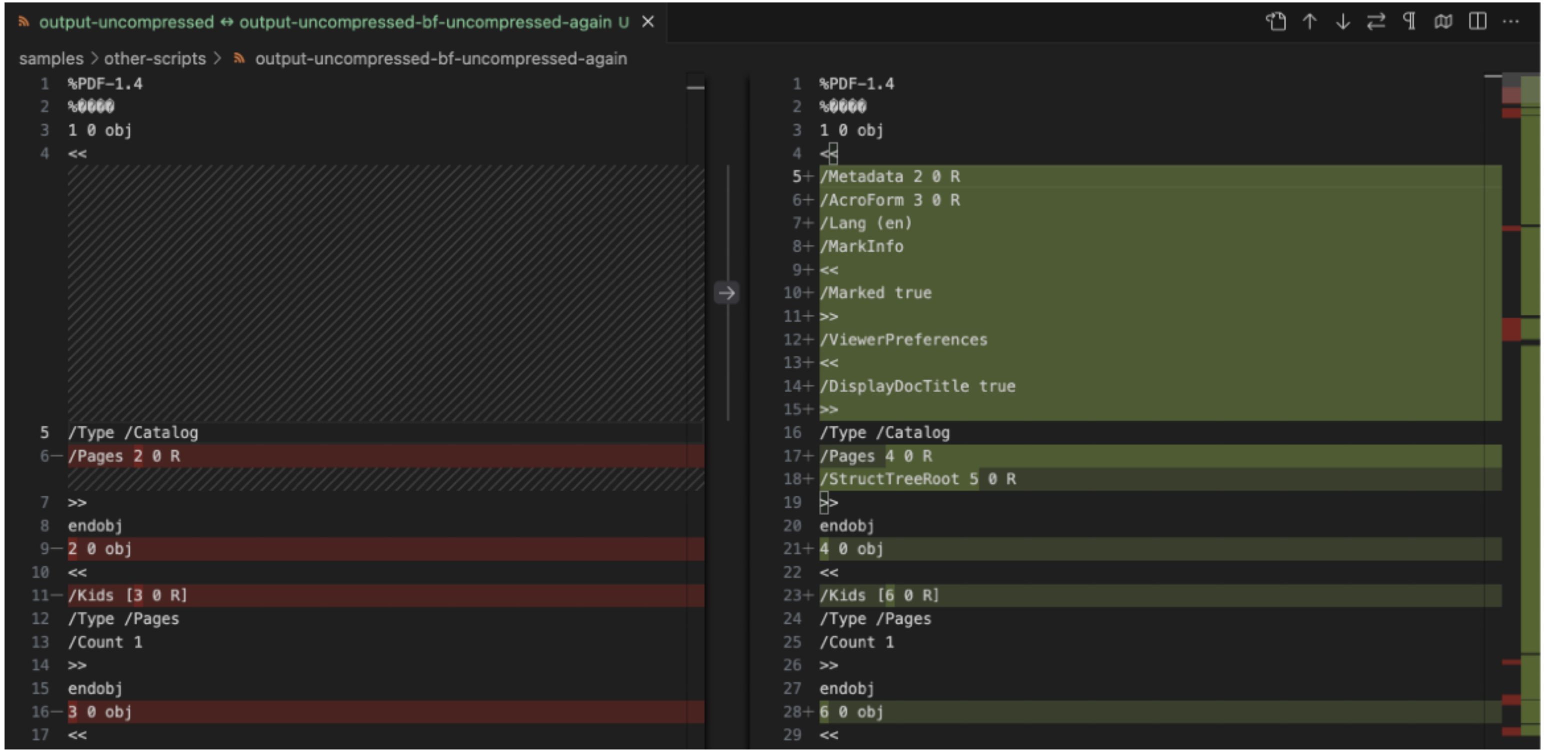
| Rule Name | Status | Description |
|---|---|---|
| Accessibility permission flag | Passed | Accessibility permission flag must be set |
| Image-only PDF | Failed | Document is not image-only PDF |
| Tagged PDF | Failed | Document is tagged PDF |
| Logical Reading Order | Needs manual check | Document structure provides a logical reading order |
| Primary language | Failed | Text language is specified |
| Title | Failed | Document title is showing in title bar |
| Bookmarks | Passed | Bookmarks are present in large documents |
| Color contrast | Needs manual check | Document has appropriate color contrast |

**Page Content**

| Rule Name | Status | Description |
|---|---|---|
| Tagged content | Failed | All page content is tagged |
| Tagged annotations | Passed | All annotations are tagged |
| Tab order | Failed | Tab order is consistent with structure order |
| Character encoding | Passed | Reliable character encoding is provided |
| Tagged multimedia | Passed | All multimedia objects are tagged |
| Screen flicker | Passed | Page will not cause screen flicker |
| Scripts | Passed | No inaccessible scripts |
| Timed responses | Passed | Page does not require timed responses |
| Navigation links | Passed | Navigation links are not repetitive |

# Core Challenge/Problem — Why is this an ML Topic?

- **Code Generation** Perspective …

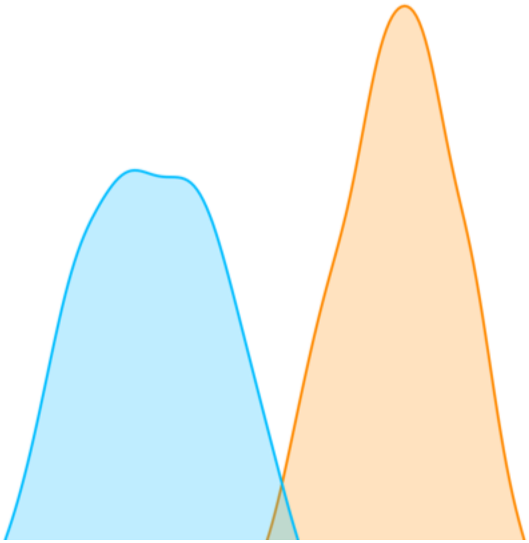


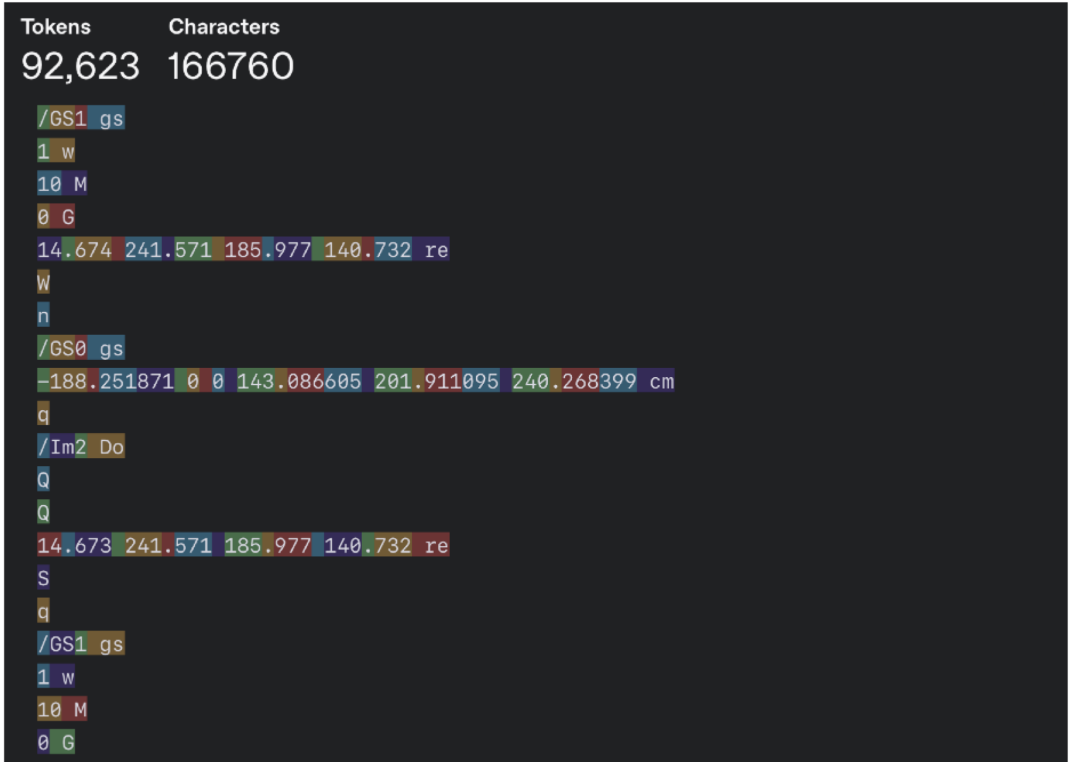
# Core Challenge/Problem — Why is this an ML Topic?

- **Code Generation** Perspective …
  - **Considering PDF file source code or structured representation**
  - not arbitrary byte sequences including binary, which is found in PDF files
- **Encoding barrier**: LLM interfaces are designed to output text, not raw binary. Raw binary tends to be interpreted as UTF-8/ASCII text, which often shows up as garbled symbols. Direct binary output is usually corrupted unless wrapped in a safe encoding (Base64, hex)
- **Tokenization limits**: LLMs don't think in bytes, but in tokens. A model can try to produce sequence of tokens that looks binary, but whether it is byte-perfect is another matter
  - For structured binaries (e.g., PDF, ZIP, ELF executable), a single incorrect byte breaks the file

# Core Challenge/Problem — Why is this an ML Topic?

- Code Generation Perspective … Preview: the **Challenge**
    - **<u>OOD</u>?**





# Core Challenge/Problem — Why is this an ML Topic?

- **Theory: Sequence-to-Sequence Task**

$$P(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{n} P(y_t \mid y_{<t}, \mathbf{x})$$

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{x})$$

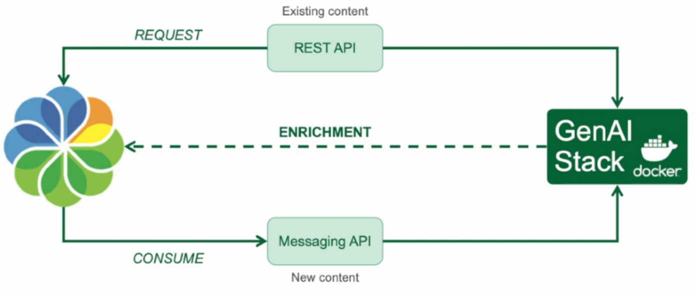
$$\mathcal{L} = -\sum_{t=1}^{n} \log P(y_t \mid y_{<t}, \mathbf{x})$$

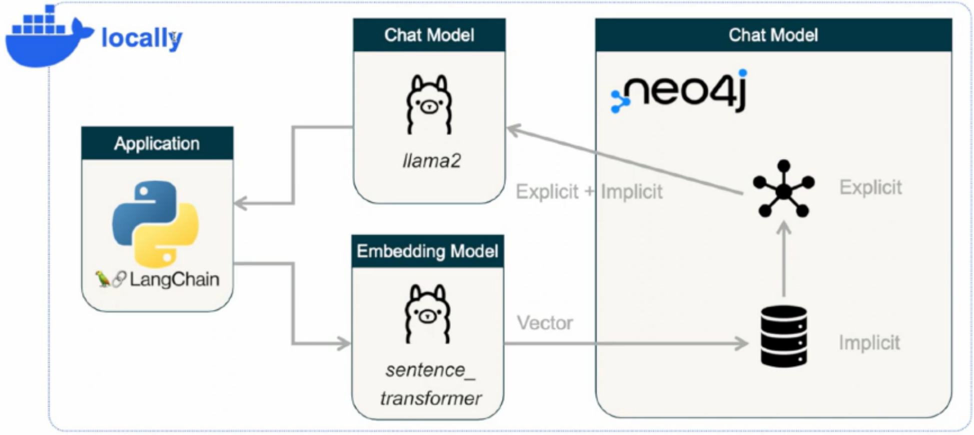
- **Models** and Prompt Engineering Analysis:
- Causal LLMs, decoder-only transformers
  - Later tests of o3/R1, “Reasoning LLMs” which introduce extra training regimes/ intermediate “scratchpad” results

| **Model** | **Observation using Prompt B.3** |
|---|---|
| Mistral (used so far) | See Table 6.1, Prompt B.3. |
| Llama3 | Comparable to Mistral - LLM does generate code but rather a high level description and interpretation of the task. |
| ChatGPT o3 via Web Client/Chat | Generates code, consulted for comparison: this is promising, finally. Elides certain content still, however, including in the output lines like `... (content truncated for brevity) ...`, for example. |
| Gemini 2.5 Pro via Web Client/Google AI Studio | Code outputs worked directly (no descriptions like Llama 3 for instance) but there was looping without reaching an end of the document code observed, in three of the five experimental runs. Examples up to error messages by the client were included in the outputs collection. Inference run-times were all above five minutes, further suggesting uncontrolled looping. |
| Llama3.3 | Similarly to Llama3, does not generate code but rather a high level description and interpretation of the task. (Not a single code file was generated even by this advanced base model, suggesting a different focus of the model and/or training data. This is not analyzed in detail as part of this work.) |
| Llama3.1 | Similar to Llama3 and 3.3. |
| DeepSeek-R1:1.5b | Similar again. |
| DeepSeek-R1(-0528):8b | Responses include thought blocks as shown in the main text, but responses are comparable to other models: no direct PDF source code is prodced in any of the five examples as it turns out. |

# The Setting and Technical Situation

- Cheaper, Big LLMs; API Vendors
- Not of Interest for this Project
  - OCR mainly
- ECM: fully on-premises vs API integration
  - Overview of the Solution Design (Practical Work)



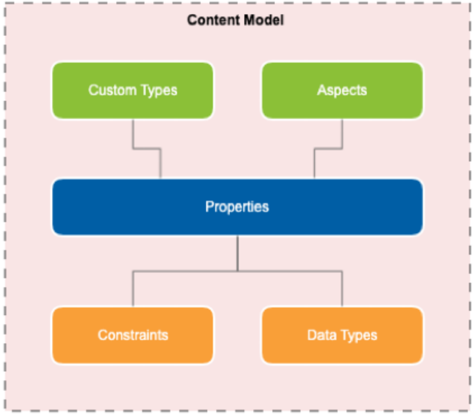

JOHANNES KEPLER UNIVERSITY LINZ

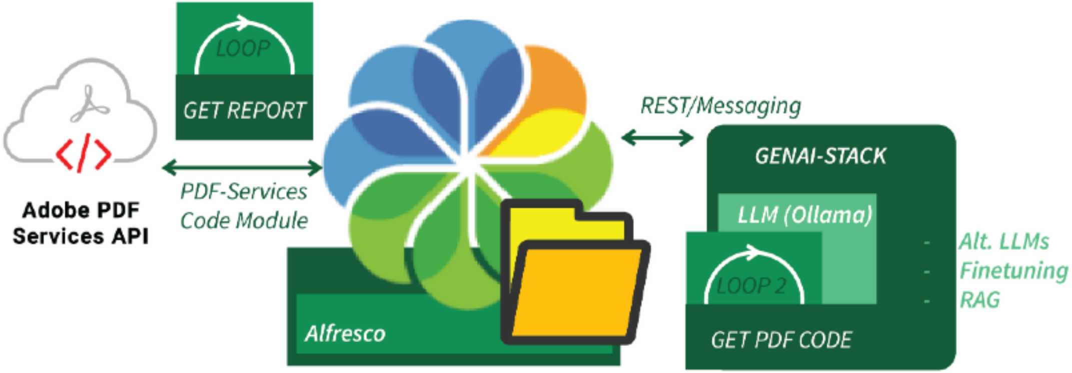
# Current Legal Context

- **European Accessibility Act** (EAA)  (Directive 2019/882/EU), building on:
  - Web Accessibility Directive (2016/2102/EU) [2a]
  - European Public Procurement Directives (2014/24/EU [2b] and 2014/25/EU [2d])
  - European Electronic Communications Code (2018/1972/EU) [2c]
- EAA explicitly applies to a wide range of products placed on the market after 28 June 2025
- On the service side, obligations cover telecommunications, audiovisual media services, passenger transport (websites, apps, e-tickets, information systems, self-service terminals), consumer banking, e-books and dedicated software, and e-commerce services
- requirements directly link to standards such as **WCAG** and **PDF/UA**
- various harmonized European standards EN for aligning products and services now
- for **ECM**: platforms evolve from passive storage and retrieval systems to active guardians of compliance for baseline of accessibility as required for regulated products and services

# ECM Implementation (Part I)

- *Brief **Demo** of Alfresco: **Demo 2***
- **Outline**: Integration into the Content Model of a simple Double-Loop LLM Call Routine, additionally prepare a Accessibility Checker Report for a meta-informational approach
- Potential value for **future work** in this domain as a relatively solid **platform**
  - Working with PDFs is intuitive and clear
  - Relevant API integrations

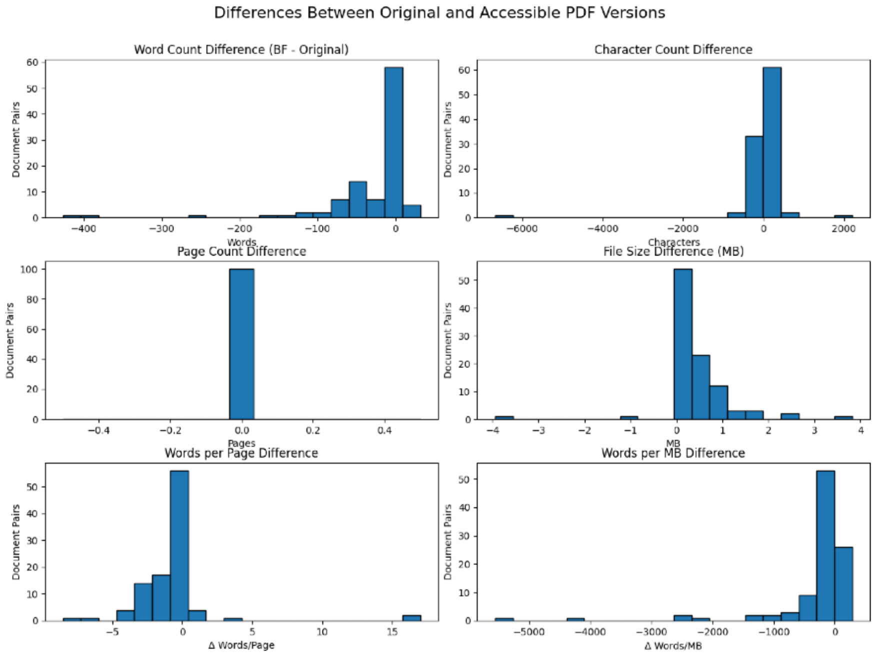


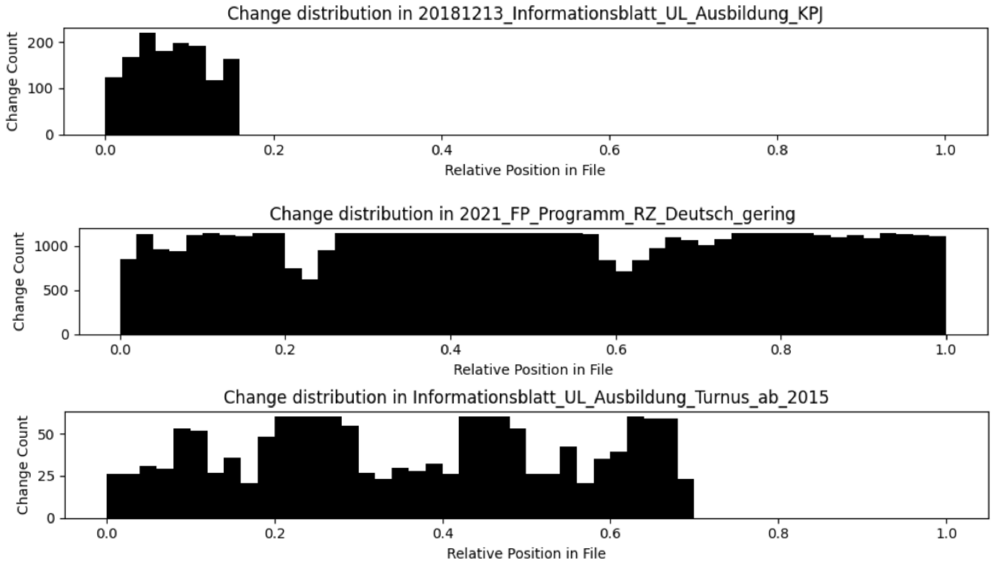



# In-context Learning, Fine-tuning and Meta-Information Approaches (Part II - Focus)

- Data: Non-accessible and accessible PDF counterparts (**Fine-tuning dataset**, 100 pairs)

Smaller **Test set** of 12 pairs is publishable




JOHANNES KEPLER UNIVERSITY LINZ

# Method: Fine-tuning

- Best prompt was used to test with different fine-tuned models
- Base-models: DeepSeek-R1-0528:8b and Llama3.1
- Tools for Fine-tuning: Unsloth (like HuggingFace Transformer Library) & Alpaca/Ollama
- Fine-tuning Approach: PEFT and LoRA

**Model and PEFT Configuration**

```
tokenizer = AutoTokenizer.from_pretrained("meta-llama/Llama-3.1-8B-Instruct")
model = FastLanguageModel.from_pretrained("unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit")
model = FastLanguageModel.get_peft_model(
    model,
    r=16,
    target_modules=["q_proj","k_proj","v_proj","o_proj","gate_proj","up_proj","down_proj"],
    lora_alpha=16,
    lora_dropout=0.0,
    bias="none",
    use_gradient_checkpointing="unsloth",
```

# Meta-Information Report-Addition, Fine-tuning with LoRA

- parameter-efficient fine-tuning (PEFT) strategy centered on Low-Rank Adaptation (LoRA) [4]

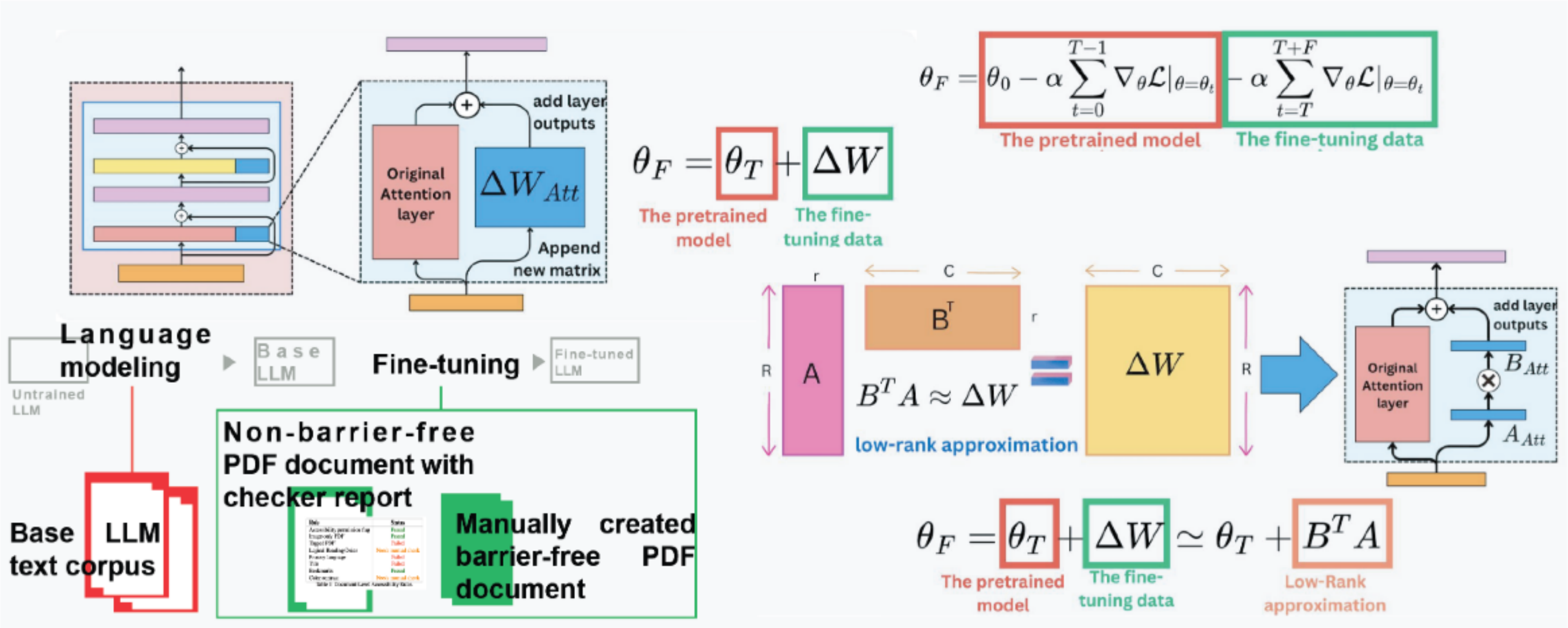



diagram based on [3] Benveniste. 2024. Understanding How LoRA Adapters Work
[4] Hu et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models

# Disadvantages of the Chosen Approaches, (Current Work & Benchmark:) Final Experiments to Improve Scores

- **Fine-tuning including accessibility reports** to test adding prompt meta-information
- Observations: OOD? Small token sizes, LLM repetition loops? Suggesting uncertainty?
- Method:
  - Load report
  - Build one bigger input string
  - Wrap into Alpaca prompt with instruction and reference
  - SFT training: loss applied on response

**Forms**

| Rule Name | Status | Description |
|---|---|---|
| Tagged form fields | Passed | All form fields are tagged |
| Field descriptions | Passed | All form fields have description |

**Alternate Text**

| Rule Name | Status | Description |
|---|---|---|
| Figures alternate text | Failed | Figures require alternate text |
| Nested alternate text | Failed | Alternate text that will never be read |
| Associated with content | Failed | Alternate text must be associated with some content |
| Hides annotation | Failed | Alternate text should not hide annotation |
| Other elements alternate text | Failed | Other elements that require alternate text |

**Tables**

| Rule Name | Status | Description |
|---|---|---|
| Rows | Failed | TR must be a child of Table, THead, TBody, or TFoot |
| TH and TD | Failed | TH and TD must be children of TR |
| Headers | Failed | Tables should have headers |
| Regularity | Failed | Tables must contain the same number of columns in each row and rows in each column |
| Summary | Failed | Tables must have a summary |

**Lists**

| Rule Name | Status | Description |
|---|---|---|
| List items | Failed | LI must be a child of L |
| Lbl and LBody | Failed | Lbl and LBody must be children of LI |

**Headings**

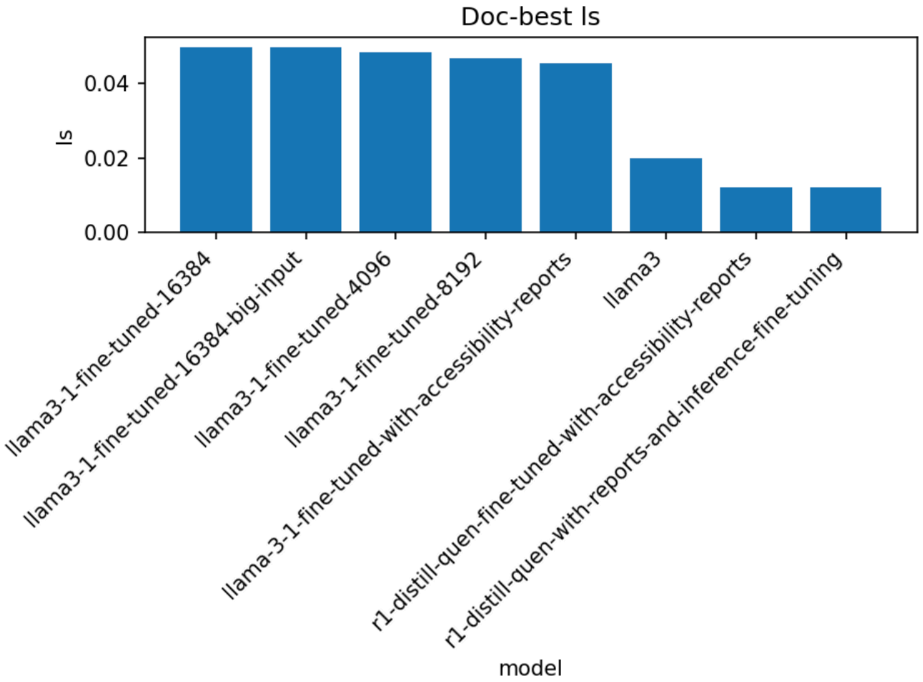
| Rule Name | Status | Description |
|---|---|---|
| Appropriate nesting | Failed | Appropriate nesting |

# Results

## With meta-info

- BLEU/ROUGE-L/METEOR are ~0–0.06 (very low)
- CER/WER are ~0.83–0.99 (very high → bad)
- Length: hyp_length_tokens is usually 5–20× smaller than ref_length_tokens (e.g., 4–611 vs 428–15k) — which is to be expected to a degree
    - Need to understand scoring between models better to get at this
    - But first: **are these model certain of what they are producing, when hypothesis/references do not matcl well in the end**?

- Similar results for this experiment

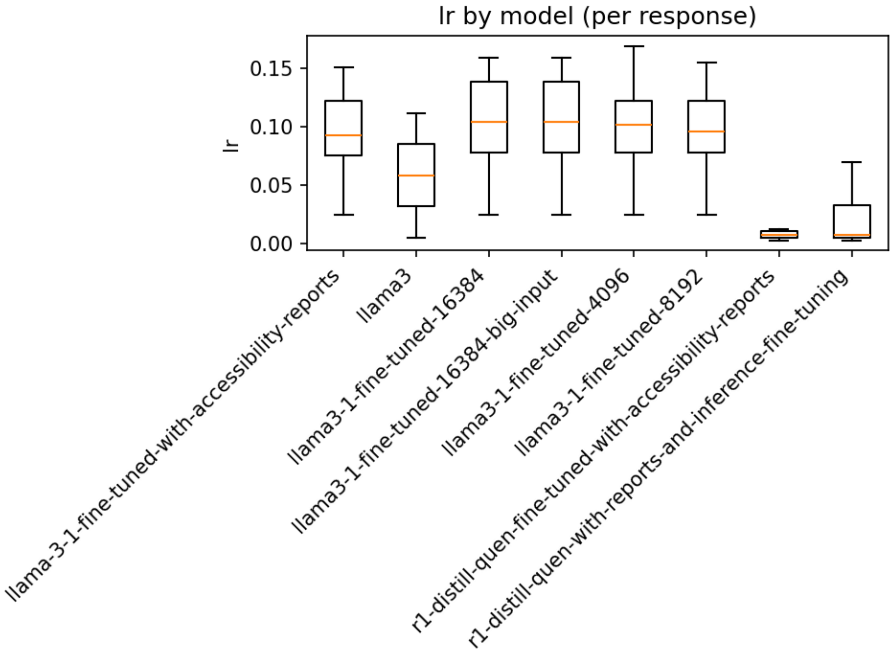


# Results

- BLEU/ROUGE-L/METEOR are ~0–0.06 (very low)
- CER/WER are ~0.83–0.99 (very high → bad)
- Length: hyp_length_tokens is usually 5–20× smaller than ref_length_tokens (e.g., 4–611 vs 428–15k) — which is to be expected to a degree
  - Need to understand scoring between models better to get at this
  - But first: **are these model certain of what they are producing, when hypothesis/references do not match well in the end**?

## With/without meta-info

- Fairly consistently similar results



# <u>OOD-Metrics</u>: Background

$$\mathcal{L} = -\frac{1}{N}\sum_{t=1}^{N}\log p(x_t \mid x_{<t}) \qquad \text{(average cross-entropy loss)}$$

$$\text{PPL} = \exp(\mathcal{L}) \qquad \text{(perplexity)}$$

- **Perplexity**: sequence-level measure derived from average negative log-likelihood
  - Lower values ⇒ model finds the sequence highly probable (more confident)
  - Higher values ⇒ model finds the sequence unlikely (more uncertain)
- **Conditional** vs. Unconditional Scoring
  - Self-likelihood: scoring the completion alone tends to be optimistic
  - Conditional perplexity: score completion while conditioning on the prompt (ignoring the prompt in the loss)
- Token-Level Uncertainty Signals
  - Predictive entropy (distribution spread over the next token) higher ⇒ more uncertainty; lower ⇒ more confidence
  - Top-1 probability: simple proxy for confidence at each step

$$H_t = -\sum_{v\in V} p(v \mid x_{<t}) \log p(v \mid x_{<t})$$

$$\bar{H} = \frac{1}{N}\sum_{t=1}^{N} H_t$$

# OOD-Metrics: Chosen Approach and Measured Perplexity

1. Generate completion (greedy for determinism or sampled for probing)
2. Compute conditional perplexity on the completion (prompt masked out of the loss)
3. Compute mean entropy and mean top-1 probability across completion steps, using next-token logits at each step

Conditional perplexity for the top-performing models so far (completion only) gets values ~1–1.5, which is **extremely low, i.e. the model is very confident**/the tokens were highly predictable — but **low perplexity ≠ good output** (we see loops, repetition, filler, not real PDF object code)

**Mode collapse/degenerate loop**: When the model falls into a repeating pattern (e.g., "BT … ET BT … ET" forever), the next token is very predictable — known to occur during fine-tuning, as the model learns to generate text that accomplishes the specific task, but loses ability to generate other forms of text.

Low perplexity reflects **predictability**, not quality: also called the likelihood trap [5]

[5] Zhang et al. 2020. Trading Off Diversity and Quality in Natural Language Generation.

# Speaking of Scores: NLP Measurements Used

- BLEU

$$\mathrm{BLEU} = \mathrm{BP}\ \exp\left(\sum_{n=1}^{4} w_n \log p_n\right) \qquad \mathrm{BP} = \begin{cases} 1, & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \leq r \end{cases}$$

- ROUGE

$$\text{ROUGE-N} = \frac{\sum_{g_n \in \mathrm{Ref}} \min(\mathrm{count}_C(g_n), \mathrm{count}_R(g_n))}{\sum_{g_n \in \mathrm{Ref}} \mathrm{count}_R(g_n)}$$

- METEOR

$$\mathrm{METEOR} = F_{mean}\,(1 - \mathrm{Pen}), \quad F_{mean} = \frac{10PR}{R + 9P}$$

# Speaking of Scores: NLP Measurements Used

- Edit Distance (Levenshtein)

$$\operatorname{lev}(a,b)=\begin{cases}|a| & \text{if } |b|=0,\\ |b| & \text{if } |a|=0,\\ \operatorname{lev}\big(\operatorname{tail}(a),\ \operatorname{tail}(b)\big) & \text{if } \operatorname{head}(a)=\operatorname{head}(b),\\ 1+\min\Big\{\operatorname{lev}\big(\operatorname{tail}(a),\ b\big),\ \operatorname{lev}\big(a,\ \operatorname{tail}(b)\big),\ \operatorname{lev}\big(\operatorname{tail}(a),\ \operatorname{tail}(b)\big)\Big\} & \text{otherwise.}\end{cases}$$

- and Combinations —
  - LS and LR were plotted so far

$$\textbf{CER}=\frac{d}{|R|},\qquad \textbf{WER}=\frac{d_w}{|R_w|}\quad \text{(edit distance at token level)},$$

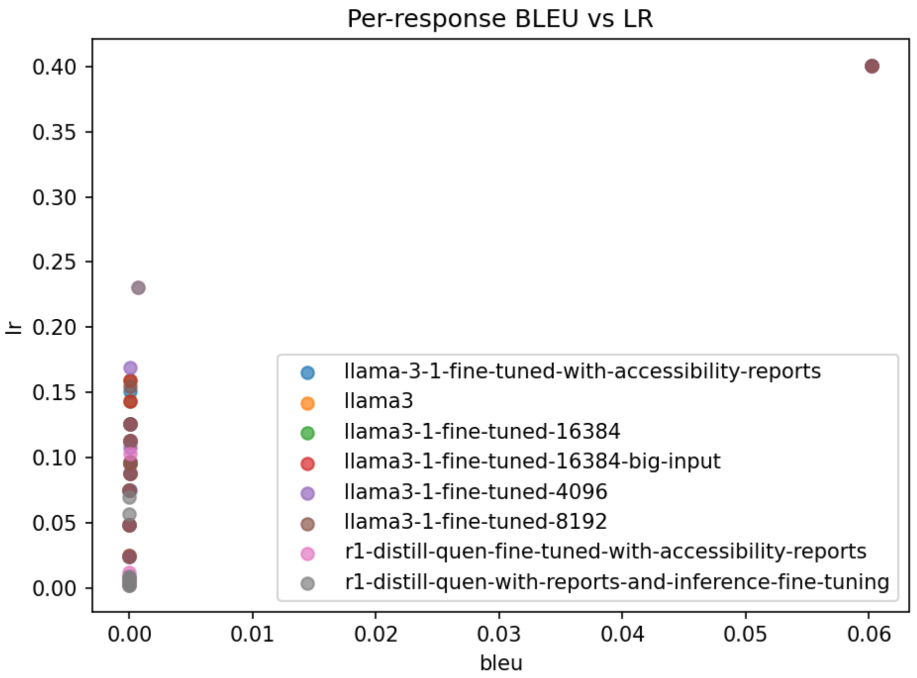
$$\textbf{Levenshtein Similarity (LS)}=1-\frac{d}{\max(|R|,|C|)}\in[0,1],$$

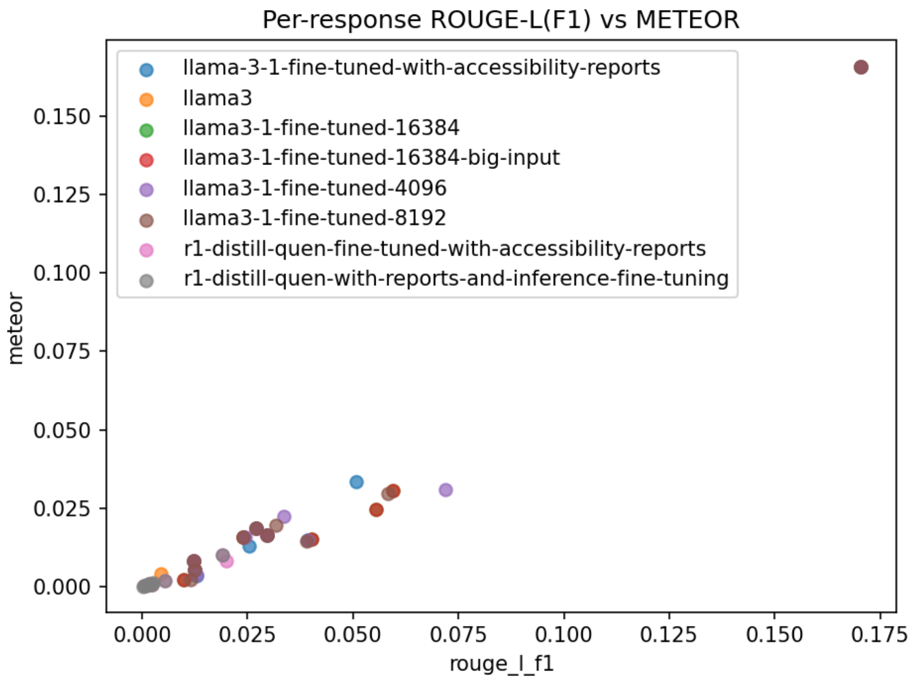
$$\textbf{Levenshtein Ratio (LR)}=\frac{|R|+|C|-d}{|R|+|C|}\in[0,1].$$

# Results

## Both With/without meta-info

- Best model here?



# Conclusion and Outlook

- **Conclusion**: positional logic and other sub-token numerical data pose an inference challenge to the chosen class of (fine-tuned) LLMs, despite low perplexity/high certainty
  - NLP-specific metrics for measuring reference similarities of the hypothesis documents were referenced to measure quality of the output, trend based on model complexity was observed, but no improvements when using meta-information in training and inference
- **Outlook**: nuanced problem with potentially large payoff, so it might be worth:
  - (Tangent:) Exploring accessibility scoring via neural network
  - Finding ways to break down the task of PDF code generation
  - Testing future or current, but more complex, models
  - Adding test document set domains like the one introduced with this work
- **Contribution**: ECM platform, basic methodology proposal, basic model testing/observations

# Summary

We considered:

- Core Challenge/Problem — Why is this an ML Topic?
- The Setting and Technical Situation
- Current Legal Context
- ECM Implementation (Part I - Not Focus)
- In-context Learning, Fine-tuning and Meta-Information Approaches (Part II - **Focus**)
- Disadvantages of the Chosen Approaches, (Current Work & Benchmark:) Final Experiments to Improve Scores
- **Results** for this work
- Speaking of Scores: NLP Measurements Used
- OOD Metrics
- Conclusion and Outlook

# References


- [1]: Klaas Posselt and Dirk Frölich. 2019. Barrierefreie PDF-Dokumente erstellen. ISBN: 978-3-86490-487-5.
- [2]
    - a) 2015. Vorschlag für eine RICHTLINIE DES EUROPÄISCHEN PARLAMENTS UND DES RATES zur Angleichung der Rechts- und Verwaltungsvorschriften der Mitgliedstaaten über die Barrierefreiheitsanforderungen für Produkte und Dienstleistungen, de. (2015). Retrieved 08/20/2025 from https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=COM%3A2015%3A615%3AFIN
    - b) 2016. Richtlinie (EU) 2016/2102 des Europäischen Parlaments und des Rates vom 26. Oktober 2016 über den barrierefreien Zugang zu den Websites und mobilen Anwendungen öffentlicher Stellen (Text von Bedeutung für den EWR ). de. (October 2016). Retrieved 08/20/2025 from http://data.europa.eu/eli/dir/2016/2102/oj/deu
    - c) [n. d.] Richtlinie (EU) 2018/1972 des Europäischen Parlaments und des Rates vom 11. Dezember 2018 über den europäischen Kodex für die elektronische Kommunikation (Neufassung) Text von Bedeutung für den EWR. de.
    - d) 2014. Richtlinie 2014/25/EU des Europäischen Parlaments und des Rates vom 26. Februar 2014 über die Vergabe von Aufträgen durch Auftraggeber im Bereich der Wasser-, Energie- und Verkehrsversorgung sowie der Postdienste und zur Aufhebung der Richtlinie 2004/17/EG Text von Bedeutung für den EWR. de. (February 2014). Retrieved 08/20/2025 from http://data.europa.eu/eli/dir/2014/25/oj/deu
- [3] Damien Benveniste. 2024. Understanding How LoRA Adapters Work! en. (November 2024). Retrieved 07/10/2025 from https://newsletter.theaiedge.io/p/understanding-how-lora-adapters-work
- [4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs]. (October 2021). doi: 10.48550/arXiv.2106.09685. Retrieved 07/04/2025 from http://arxiv.org/abs/2106.09685
- [5] Zhang et al. 2020. Trading Off Diversity and Quality in Natural Language Generation. https://arxiv.org/abs/2004.10450