Disclaimer - read this first:
https://www.wu.ac.at/dpkm/topics/how-to-write-a-thesis/

How to write a thesis

how to get started, some Do's and Don'ts

… particularly in case I'm your supervisor ;-)

Axel Polleres

version: 2022-04-26

Some tips:

- Your thesis is (typically) a **monograph**,* it shall have a clear structure and a running **story** that you tell throughout.

- Your thesis is **not** a collection of excerpts of the literature, i.e.:
  - it is ok to start with summarizing even copying text from papers you find as a part of your structured literature review BUT:
    - Summarize what you **UNDERSTOOD**, not what you read.
  - i.e. doing excerpts is not enough, but only a first step…
  - … you rather should do a structured literature review.

* for PhD theses and Habilitation theses ***sometimes*** *also cumulative theses (i.e. collections of own original works) are allowed, but that does not apply to Master or Bachelor theses.*

Another word on literature References (especially important for the proposal)…

- …. Again, do **NEVER** just summarize text and put the reference in the end!

- References have two purposes only:
  - Pointing to places where the reader finds more/deeper explanations than you provide in your written text
  - Providing a source of where you took inspirations/your understanding for writing ***your own text*** from.

Implications:

- Use only references you have **read & understood**
- For each reference, your text should explain:
  - why you used it
  - why it is relevant for your thesis

Plus, you should be able to explain this for each reference in your own words, when I ask you ;-)

"Phases" of doing a thesis:

- The second point of the last slide cannot be stressed enough…
  - … **write-up** of your thesis should be considered a separate task/phase of your thesis project!
    - it takes considerable time
    - "mingling" this with the data/literature collection phase, i.e. collecting information and writing the thesis in one go is a common "recipe for failure"
- Consider your thesis a **project** and plan it like a project!
    - Example phases of a thesis project:
      1. **research/literature collection phase** - understand the problem
      2. Think of your **contribution – what should be the (ideal) outcome?**
      3. **methodological familiarization phase**: practice, familiarize yourself with needed technology (document what you learn here, this will be valuable for a "preliminaries section" in your thesis)
      4. **design of your approach**
      5. **implementation phase**
      6. **evaluation phase** (again, this may need a design phase for the evaluation, deciding on data to be used, how you want to analyze this data)
      7. **write-up phase**
         - start from an empty document
         - continue with a coherent structure
         - write each section from scratch, drawing from the documentation and texts you produced in the earlier phases
         - write the conclusions last, think about limitations and things you had to leave open in your work

What is your own *contribution?*

- We expect an **own contribution**, in terms of e.g.
  - **Systematically structuring** and comparing existing solutions on a problem:
    - *implement your own approach/artifact "**design science**"– and think about how to evaluate it*
    - ***reprocucability study*** (re-apply/re-implement/extend existing approches and compare your results to published ones
    - define/appply criteria, systematically "slot" your research results into these criteria - ***qualitative comparison***
    - *collect opinions of others (study/survey/interview-based)*
    - *…*

    - *i.e. it is not enough to summarize existing work sequentially "in your own words".*

    → The choice of contribution will determine your methodology! Not the other way around ;-)

The importance of **reading**:

- Start reading now…
  - Read a lot of scientific articles
  - Look at other theses and how they are structured
- Reading will help you
  - Understanding how scientific work is best structured and presented → follow the examples, but find your own "mix".
  - Read articles several times or iteratively
    - 1st time "skim"
      - Start with getting the ideas, collecting articles,
      - define which articles are in scope and which ones are "future work"
      - which ones are relevant, which ones are not
    - Read the in-scope articles in more detail
    - Be ready to read more along the way
    - Decide where you stop collecting more literature
- I.e. reading and summarizing papers helps you to
  - scope & understand

  the relevant literature

Tools you can use:

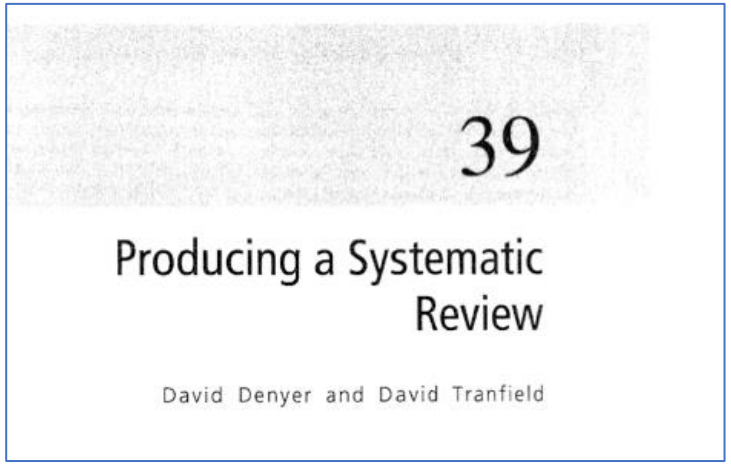
- make a mindmap
- categorize papers (post-its, labels, keywords)

A structured literature review:

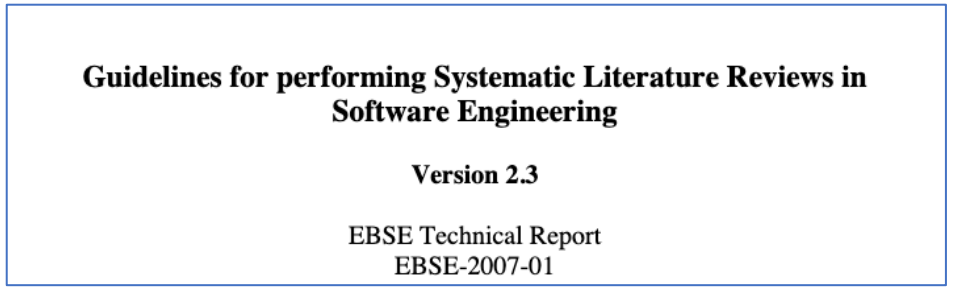
- There are techniques for collecting and scoping literature in a structured manner:
- A complete and well-done structured literature review ***may*** be
  - a part of your work to scope related works
  - worth a BSc or even an MSc thesis alone *(exception)*,

    if done exhaustively and rigorously

- First step: collect ALL references you come across in a BibTex database, and keep the bibtex entries tidy (always add authors, title, proper information how the source has been published, etc.)



https://www.cebma.org/wp-content/uploads/Denyer-Tranfield-Producing-a-Systematic-Review.pdf



https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf

https://twitter.com/moduloone/status/1449523528303198213

Some words on ***self-containedness*** … What do you need to explain? What not?

- Your thesis should be ***self-contained***, in the sense that it is readable for your colleagues, or clear where they find the explanations they need.
- I.e., someone who studied with you would be able to understand it, without having to read/redo your work, and could potentially continue your work.
  - only write things **you understood**
    - in a way that someone else can also understand them
  - write down how you understood them, i.e. either
    - explain them in your own words
    - provide a reference, where there's a good explanation that helped you understand
  - most importantly, don't copy or use terms and definitions you cannot explain.

  - Read this: https://www.inc.com/glenn-leibowitz/the-single-reason-why-people-cant-write-according-.html

Scope & Management

- One of your main tasks in regular meetings with your supervisor is to **manage the <u>scope</u>** of your thesis!

- We will push potentially *many* ideas and suggestions to you how and where you can start and which particular aspects of your thesis topic you could investigate.

- It is *your task* to define a scoped topic out of these suggestions

- i.e., take the literature tips of your supervisor as starting points, but don't follow all routes at the same time --> it's ***up to you*** to define and argue the scope and to demonstrate that you have followed the literature deeply enough.

Scope & Management

- The other main tasks in regular meetings with your supervisor is to **manage the <u>progress</u>** of your thesis!

  - Keep record of meetings (minutes) and send those to the supervisor
  - Have a time plan & milestones and discuss it with your supervisor
  - Agree on next meeting in the end of each meeting
  - Agree on concrete things you plan to achieve until the next meeting

- Do not expect your supervisor to remember the details of your thesis/last meeting:
  - i.e., start each meeting with a summary and status report and what was agreed in the last meeting!

Format:

- In English (German *really* an exception only)
  - English is the main language of our scientific discipline (and literature)
  - most terminology doesn't translate well anyway to German
  - makes your thesis more accessible internationally

- Take care of proofreading and English language check, plan time for it

- A bachelor thesis has a page limit of 40 pages text (not including cover, table of content, references, appendices).
- A master thesis has a page limit of 80 pages text (not including cover, table of content, references, appendices).
- A MBA thesis has a page limit of about 60 pages text and slightly different formatting guide by ExAc

- If you have more material, think of moving parts to an appendix.
- If you have less material, do NOT fill up pages, it is really NOT about pages.

- If you want to write an academic thesis well: ***read a lot of scientific articles***! ***start now!***

How to write a good research paper/proposal/thesis?

Make your thesis an interesting read!
i.e., don't annoy the reviewer/reader ;-)

Heilmeier‘s catchism:

George H. Heilmeier, Director of DARPA (1975 – 1977)

Critical questions for ANY research project/paper/thesis:

paraphrasing:
What is the problem?
Why is it a problem?
Why should the reader care?
What's your solution/contribution?
Hint:thesis **introduction** section should answer those!

A set of questions credited to George Heilmeier that anyone proposing a research project or product development effort should be able to answer.[1]

- **What are you trying to do**? Articulate your objectives using absolutely no jargon.
- **How is it done today**, and what are the limits of current practice?
- **What's new** in your approach and why do you think it will be successful?
- **Who cares**?
- If you're successful, **what difference will it make**?
- What are the **risks** and the payoffs?
- How much will it **cost**?
- **How long** will it take?
- What are the midterm and final "exams" to **check for success**?

[1] G. Heilmeier, "Some Reflections on Innovation and Invention," Founders Award Lecture, National Academy of Engineering, Washington, D.C., Sept. 1992.

Structure:*

My own „default“ structure of a research paper (mainly suitable for the "case study" style):

0. Abstract
1. Introduction (Motivation)
2. Background
   - Introduce background as necessary for the reader
   - Often useful to add a „running example/scenario“
3. <Your proposed solution>
4. Evaluation
5. Related works
6. Future Work & Conclusions
7. References

- (Appendices: proofs, details on experiments)

Important:
DON'T USE *THIS (or any other "generic"* structure!!!

Your supervisor (and other readers) should already see from the content structure what you worked on!!!

→ Variations of this scheme may apply, depending on your **audience & purpose**:

- Purpose: magazine article, vs. scientific journal vs. Seminar article vs. Diploma thesis)
- Audience: research community (Database vs. Economists), students vs. experienced researchers vs. Layman audience
- The main thing is : A thesis is a monography → Tell **one coherent story**
- Tell **Your** story: i.e. follow the structure in spirit, but adapt it to your needs, do not blindly use the same section headings!

***) For PMBA theses:** *see also Section 2.2 of the PMBA Thesis guide*

Structure: "award-winning" example" (TALENTA)

Bachelor Thesis

Open Dataset Archive

Scalable dataset crawling with efficient archiving and the investigation of changes between versions.[1]

Contents

1 Introduction **1**
1.1 Problem Overview . . . 3
1.1.1 Data Type Detection . . . 3
1.1.2 Detection of Changes . . . 4
1.1.3 The Storage-Recreation Trade-off . . . 4
1.1.4 Workload-management and Scalability . . . 4
1.2 Research Question . . . 4
1.3 Thesis Structure . . . 5

2 Preliminaries & Background Literature **5**
2.1 Data Types . . . 5
2.1.1 Unstructured Data . . . 6
2.1.2 Semi-structured Data . . . 6
2.1.3 Structured Data . . . 6
2.2 Architectural Hurdles . . . 6
2.2.1 Host Politeness . . . 6
2.2.2 Dynamic Crawl-Rate . . . 7
2.2.3 Scalability . . . 7
2.3 Related works on data archiving and versioning . . . 8
2.3.1 Online Platforms . . . 8
2.3.2 Git and SVN . . . 9
2.4 Preliminaries and Technologies used in this thesis . . . 10
2.4.1 Databases . . . 10
2.4.2 Programming Languages and Concurrency . . . 11
2.4.3 Kubernetes, NGINX Ingress and Reverse Proxy . . . 12

3 Requirements and Services **12**
3.1 Primary Requirements . . . 13
3.1.1 Application Programming Interface . . . 14
3.2 Secondary Requirements . . . 14

4 Implementation **15**
4.1 Architecture . . . 16
4.1.1 System Structure . . . 17
4.1.2 Sequence Diagram . . . 18
4.1.3 Database Model . . . 19
4.2 Data Access & Client Interface . . . 20
4.2.1 Public API . . . 20
4.2.2 Private API . . . 21
4.2.3 SPARQL Endpoint . . . 21
4.3 Traffic and Workload-management . . . 23
4.3.1 Parallelization and Scalability . . . 23
4.3.2 Dynamic crawl frequency . . . 24
4.4 Data Management . . . 25
4.4.1 Type Detection and Data analysis . . . 25
4.4.2 Compression . . . 26
4.4.3 Resource Handling . . . 26
4.5 Dependencies and Open Issues . . . 27

5 Findings **28**
5.1 Corpus of the archivers database . . . 28
5.2 Monitoring and Bench-marking . . . 29

6 Conclusion and Further Research **31**

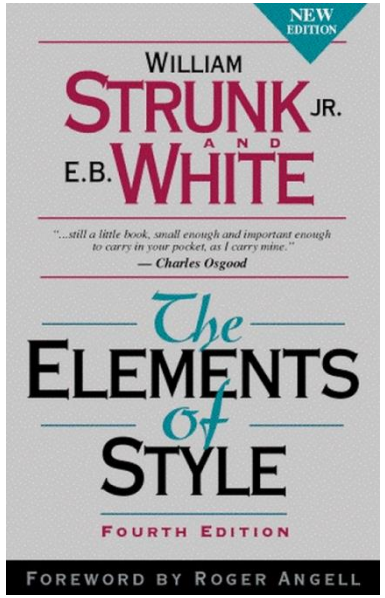
7 Acknowledgements **33**

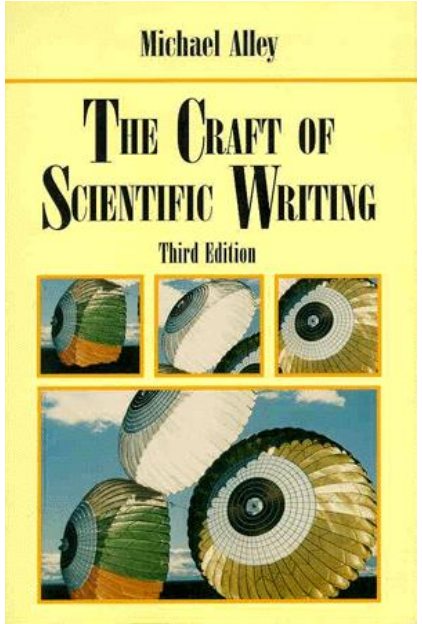
Some words on style:

- use passive voice sparsely (rather use an inclusive "we")
- don't state the obvious
- Don't use subheadings directly after another heading

…

There are tons of books on Style…
… some examples from my shelf



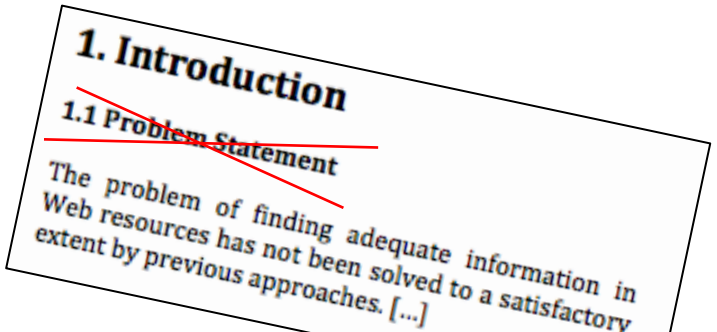


If you look for good examples, take the base articles as a reference!

Some words on style

- Most importantly: Don‘t annoy the reader/reviewer
  - **use a spell-checker** ;-)
  - avoid to repeat yourself
  - explain the “big picture” before you jump into technical details
  - avoid to state the obvious
  - Avoid “Germanisms”, e.g.
    - again: in English ***passive voice*** is less common than in German
    - In English, often people avoid using ”I” in favor of a (reader-inclusive) “we”, even in single author papers!
  - use proper citation
  - be concise

    „If I had more time I would have written a shorter letter“ *(Blaise Pascal? Mark Twain?)*
  - don‘t jump, follow a consistent line of arguments („roter Faden“)
  - always **try to consider the viewpoint of your reader/audience**
- **Don't don't** (i.e. avoid short forms like don't doesn't, can't …)
- Use section structuring properly, e.g.


1. Introduction

1.1 Problem Statement

The problem of finding adequate information in Web resources has not been solved to a satisfactory extent by previous approaches. [...]

More **bad** examples 1/4:

- Avoid to state the obvious:

~~„This seminar paper was written for the course `Forschungsseminar Systemanalyse` in Summer term 2015“~~

... Not relevant. Look at accepted conference papers, nobody writes:

~~„This paper was written as a submission to the World Wide Web conference.“~~

i.e. the motivation/introduction section should explain **what is the problem**, not why you had to write the thesis. ;-)

More **bad** examples 2/4:

- use proper citation in the text, e.g. do not use citation as subject of a sentence:

~~[5] developed the Web of concepts approach by …~~

- Better, e.g. :

Dalvi et al. developed an approach called ``Web of concepts" [5] …

or:

The ``Web of concepts" approach [5] …

References

…

[5] Nilesh Dalvi, Ravi Kumar, Bo Pang, Raghu Ramakrishnan, Andrew Tomkins, Philip Bohannon, Sathiya Keerthi, and Srujana Merugu. A web of concepts. *Procceedings of ACM PODS 2009*, pages 1–12, 2009.

…

More **bad** examples 3/4:

- use of domain-specific terminology in your paper without explaining them, or giving a concrete reference where it is explained. E.g. if you write…

> ~~*"The **k-means** algorithm always finds a solution, which is not necessarily the **optimal** solution. Finding an optimal partitioning belongs to the class of **NP-hard** problems."*~~

- … What do you mean by ***optimal***? Can you define it? Can you expect the reader to know what ***NP-hardness*** means? Would you be able to explain it, if your supervisor asks you? If not:
  - Read up on it!!!!
  - Explain it in the preliminaries

- Better:
  - Explain **all** relevant domain-specific terms in a leading "Background" or "Preliminaries section" and/or add a good reference *(**ATTENTION**: only explain those terms relevant for your theses, finding the right balance between explaining too little and too much is an art ;-))*

More **bad** examples 4/4:

- similarly, use of abbreviations without explaining them:

~~*"We want to show which sub-question can be answered using SPARQL queries on the MAKG."*~~

- Better:
  - Explain each abbreviation at its first use in the thesis, provide references:

"We want to show which sub-question can be answered using queries in the SPARQL query language [1] on the Microsoft Academic Knowledge Graph (MAKG).[1]"

---

[1] *available at http://ma-graph.org/, last acceessed 27.2.2021*

Follow back to **original** References:

"We use PageRank [1], … "

- Cite the original, e.g.
  - don't cite Wikipedia for PageRank

    ~~[1] Wikipedia "PageRank" . https://en.wikipedia.org/wiki/PageRank, last accessed 27.2.2021~~

  - don't cite another paper that uses PageRank, e.g.

    ~~[1] Ljosland, Mildrid. "Evaluation of Web search engines and the search for better ranking algorithms." *SIGIR99 Workshop on Evaluation of Web retrieval*. 1999.~~

  - but cite the original reference:

    [1] Brin, S.; Page, L. (1998). "The anatomy of a large-scale hypertextual Web search engine" (PDF). Computer Networks and ISDN Systems. 30 (1–7): 107–117. doi:10.1016/S0169-7552(98)00110-X. ISSN 0169-7552.

- General rule: cite ***webpages*** **ONLY** as a last resort (if there's no good textbook or academic paper explaining it), and, for Webpages, always add a "last accessed" date.
  - Additionally, Check if URLs are indexed at the Web archive and if yes , ideally refer to the specific version you mean, e.g. the example from the last slide:
    - https://web.archive.org/web/20201206004707/http://ma-graph.org/ refers to the last indexed version on the Web archive from 06 December 2020.

- … I could go on forever here, but once again to conclude:
  - learn from reading good papers!
  - not only from their content, but also from their
    - structure
    - use of references
    - explanations of related backgrounds
    - style
  - start reading and searching for literature now! ☺

Examples of good theses?

- Some excellent theses that truly fullfilled or exceeded these expectations (at various levels, incl. 5 WU TALENTA awards!):

http://polleres.net/supervised_theses/

https://semantic-systems.org/student-thesis/

(Hint: unless you're a PhD student …

… of course you're NOT expected to do a CS PhD thesis ;-)))

To get equally great results:

- **plan & manage** your thesis (and your supervisor ;-))

Further information

- formatting
- Why use LaTeX?
  - the main tool used in scientific writing!
    - Makes keeping of references **really** easy (using BibTex)!
    - For our undergraduate thesis we actually provide a LaTeX thesis template it (which in case someone is interested, we could adapt):
    - You can work using online tools, e.g. https://www.overleaf.com/ (many good tutorials on their page as well!

- … again: find it here: https://www.wu.ac.at/dpkm/topics/how-to-write-a-thesis/

Code? Data?

- If you produce code or data for your thesis, **wherever possible**:
  - Be ***FAIR*** (cf. https://www.go-fair.org/fair-principles/):
    - Findable
    - Accessible
    - Interoperable
    - Reusable
  - Apply-best practices:
    - E.g.
    - use a versioning tool/repository
    - use a git repository (we can provide one at the institute)
    - Add a licences for reuse (ideally an **Open** License!)

  - i.e. make your work re-usable for other students who could build up on it!

A word on plagiarism…

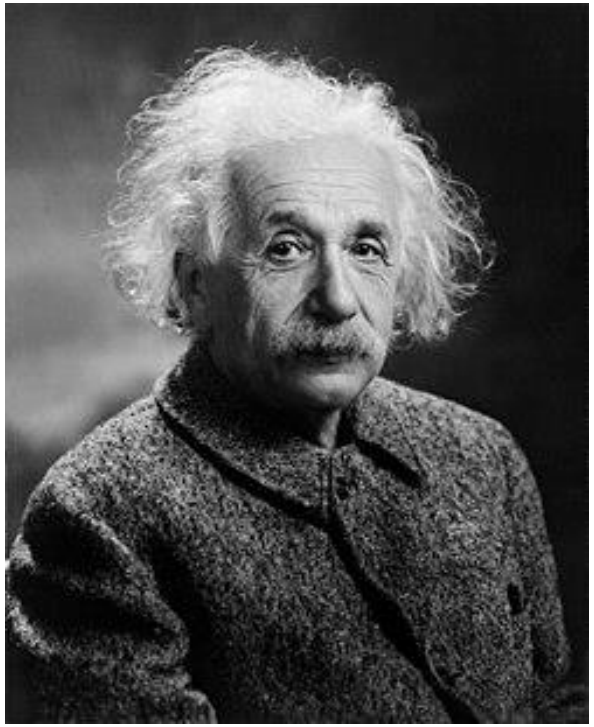
- ALL (no exceptions) inspirations you took from elsewhere should be referenced

- We us e a plagiarism checker!
  - reformulate and explain things you read elsewhere IN YOUR OWN WORDS
  - any literal quotes have to be clearly marked as such (*using quotes and italic*)

- Also **figures and images** need to be referenced!!!!
  - I also use image search engines to check figures

- Again: follow references to the **original** source! (cf. slide 20)

A word on plagiarism(?)… LLMS



Before asking an AI model for help, think twice:

*”If **you** can't explain something **simply**,*
* **you** don't understand it well enough”*

- Albert Einstein

A thesis (and its defense) shall first and foremost proof that YOU have understood the subject matter:

Implication: this is why we need to orally exam you about your thesis, ensure that you can explain it, and that you understand it.... *to the level of every explaining and arguing any sentence and line of code* you wrote. That's the very learning goal of a thesis.

No “Slop grenades”!

https://en.wikipedia.org/wiki/AI_slop

Interact with your supervisor, rather than with AI tools only!

p

AI Overview

A **"slop grenade"** is a slang term for unedited, bulk generative AI content (such as text, emails, code, or documents) that is generated and passed on to colleagues or recipients without review, verification, or added human value. LinkedIn · Jodie Cook +2

A word on plagiarism(?)… LLMS

- Large-Language models like GPT, Claude, Gemini LLAMA, and friends
  - are like “pocket calculators”
  - but (still!) make mistakes (hallucination)
- In general, we discourage their use for a thesis: you should learn how to write a good text/thesis yourself**!**
- **Agree about AI usage bounds with your supervisor <u>as part of the proposal</u>**
- Don’t take any output of an LLM into your thesis unreflected/unchecked
- **If** you use an LLM, keep/provide a **transcript** of your interactions!
- **NEVER** feed *sensitive*, *personal, or copyrighted* information into an LLM!
- Be **ALWAYS** able to explain anything you write yourself!
- **Summarize your own contribution** in your thesis, e.g. for a survey, how you structured the literature, which criteria you defined.

Sample guidline from a syllabus… apply analogously! ;-)

*“**AI Policy**: In recent years, generative AI technologies such as GPT and Github Copilot have become proficient in writing certain types of code and can help programmers increase [illegible]ductivity. However, the developers of these tools have cautioned that for novice programmers, [illegible] these tools may result in over-reliance and a worse lea[illegible]g ou[illegible] this cla[illegible]You sh[illegible] this wo[illegible] help [illegible]ore app[illegible]priate[illegible] and closed-com[illegible] ts if you[illegible]e used [illegible]*

FAIR AI usage:
(1) keep track of all thesis related use in an obligatory **List of aids (Hilfsmittelverzeichnis).**
(2) eparate use for your thesis from private use !!!
<u>(keep separate chat/agent instruction logs)</u>

Any additional questions?

In this order:

- contact your supervisor!
  - write a friendly reminder if you don’t get an answer within 2-3 days
- backup (sometimes emails may get lost, just too many ☹):
  - contact backoffice@ai.wu.ac.at