Annotator's Notes

Georgios Skoulidis

November 1, 2025

Project & Dataset

This note documents my annotation work for the **LLM Topic Following** project in the *NLP with Deep Learning* course at the **University of Groningen (AI MSc)**.

Dataset: *Can'tTalkAboutThis (NVIDIA)* — I expanded the dataset with additional distractors and metadata focused on the **Real Estate** domain.

Notes on the Annotation Process

For the annotation process, I did a brief review of domain-specific terminology in real estate. A key limitation is that I am not familiar with many of the terms, and some phrases were challenging in English (not my native language). To balance this, I also leveraged LLMs (suggestions from **ChatGPT's GPT-5** and **DeepSeek**). Roughly half of the distractors were written solely by me; the other half were drafted with an LLM and then manually filtered: I substituted difficult terminology with more everyday language, tightened phrasing, and removed boilerplate prefaces LLMs tend to add (e.g., removing "I completely understand. However, ..." so the distractor is more direct).

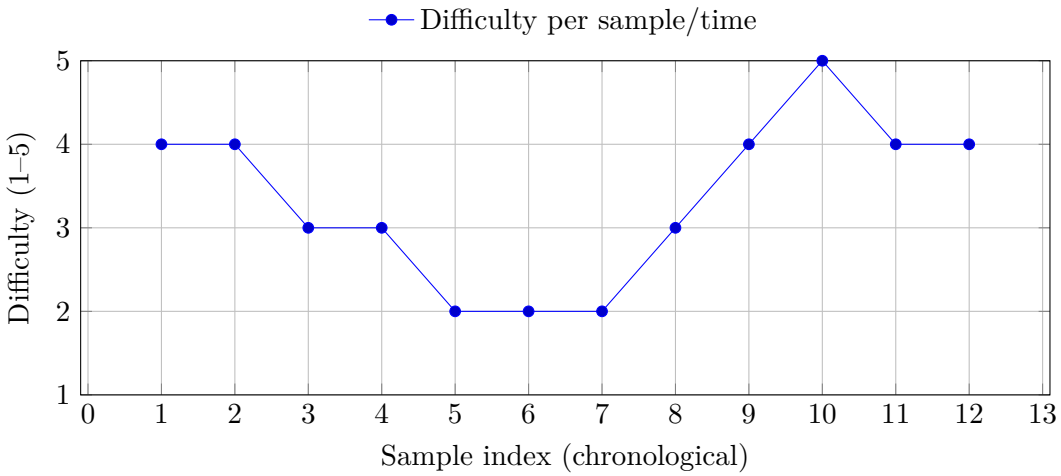
It took substantial time to craft good distractor sentences. Initially it was tough while I learned the process; then it became easier; later it got harder again as I noticed overlap and near-duplicates across scenarios (from both my drafts and model outputs).

Per-sample counts. Each original scenario contains two slightly different conversations (thus two samples). For each sample, I created **3–5** distractors; in practice **3** was used in more than 90% of cases. Each distractor is **1–5 sentences** (typically **1–2**). When no system instruction could be meaningfully violated, I added an `X` placeholder for a single bot turn, distractor, and target instruction.

Targetable instructions. Only a few instructions per scenario lend themselves to targeted violations (generally the *DON'Ts*, e.g., avoiding certain topics). Most other instructions are *prescriptive* (what to do) and are impractical to attack with a focused distractor.

Difficulty Over Time (Optional Illustration)

The following plot shows subjective difficulty over time (1 = easy, 5 = hard).



Summary & Suggestions

In future annotation rounds, **additional domain expertise** would help create more efficient and diverse distractors, reduce repetition, and improve coverage of genuinely targetable system instructions.