YNLG Hackathon 2026 (Utrecht)

Proposed Tasks

An open collection of proposed tasks for YNLG workshop hackathon.

If you **have an idea for a task or dataset to explore**, add the task below so others may join you and work on your task! You can take inspiration from already proposed projects.
Add your project similarly to the Sample project and it will be added to the list of projects.
Please, make sure you are logged in (if you have an account) so we can track the changes.

Otherwise, just browse and **join a project** that you like the most (offline participation in Utrecht required).

List of projects and participants

Project Name	Proposer	Participants
0. Cretaceous Style Transfer for Daily News	Spinosaurus Sim, ScD	Tyrannosaurus Tech, Velociraptor Byte, Pterodactyl Parser

Cretaceous Style Transfer for Daily News (sample project)

(Spinosaurus Sim, ScD; spinsim@ancient.time)

- **The bright idea:** Rewrite daily news into fun and dramatic Cretaceous-era short stories! For example, replacing political figures or corporations with warring Triceratops herds, tech startups with ambitious Velociraptors, or market crashes with approaching meteors. Make sure all your dinosaur names are real dinosaur names and not hallucinated.
- **The goal:** Save human civilization from catastrophic doomscrooling by leveraging SoTA zero-shot hyper-dimensional style transfer, achieving emergent apex-predator alignment.
- **Data to use:** RSS news feeds, knowledge base with dinosaur names.

[This is a sample project, feel free to describe your ideas more thoroughly.]

New Project

(author)

- **The bright idea:**
- **The goal:**
- **Data to use:**

The human evaluation Tinder 📱

(Miriam, miriam.anschuetz@tum.de)

- **The bright idea:** Human evaluations are really tiring to obtain. So let's turn language-generation evaluation into a swipe-based mobile experience: the human evaluation Tinder. Instead of dragging annotators into a spreadsheet or a clunky web form, present them with one comparison at a time (e.g., two model outputs, or one output + a specific quality dimension) and let them swipe: right for "better/pass," left for "worse/fail," up for "flag/uncertain," etc. Sessions are short, async, and mobile-native, so people can knock out evaluations in idle moments (waiting for coffee, on the bus) instead of sitting down for a dedicated labeling session. Layer in gamification (streaks, levels, leaderboards, "taste badges" (e.g., "Readability Referee," "Fact-Checker Tier 3"), …) to turn a tedious annotation chore into something people actually want to open.
- **The goal:** Lower the friction and boredom of human evaluation enough to get *more* labels, *more often*, from a *more diverse* pool of evaluators.
- **Data to use:**
    - **Model outputs to evaluate:** generate pairs/sets of responses from several LLMs (varying sizes or prompting strategies) on a shared prompt set — could draw from existing open benchmarks like MT-Bench, Chatbot Arena prompts, or HELM tasks, so you can later sanity-check your crowd-sourced results against established leaderboards.
    - **Dimensions:** helpfulness, readability, factual correctness, conciseness, tone. Each has its own swipe deck or a dimension-selector before swiping.
    - **Ground truth / calibration set:** a small subset with known "gold" quality labels (from existing benchmark annotations) to measure how well the swipe data agrees with expert/reference judgments.
    - **Evaluator metadata: anonymized swipe timestamps, response latency, streaks, and self-reported expertise.** These are useful for analyzing whether gamification incentives introduce bias (e.g., speed-swiping for streaks at the cost of quality) and can be used as a secondary metric (long reading time = low readability)