

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).

[Cretaceous Style Transfer for Daily News \(sample project\)](#)

[The human evaluation Tinder](#) 📱

[The Classical Pipeline Strikes Back](#)

[The Qualitative Vibe Check](#)

[Who Wrote This? Stylometric Fingerprinting in Story Continuation](#)

[Joke fight](#)

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
1. The human evaluation Tinder	miriam.anschuetz@tum.de	
2. The Classical Pipeline Strikes Back		Johanna Rockstroh
3. The Qualitative Vibe Check Aka: Steam dataset oldschool vs. vibe	onderkova@ufal.mff.cuni.cz	Aleksandr Semenikhin Joseph Moretto
4. Who Wrote This? Stylometric Fingerprinting in Story Continuation	yongxin.zhou@univ-grenoble-alpes.fr	Menan Velayuthan (m.velayuthan@uu.nl) David Liu (david.liu2@unsw.edu.au) Leruo Sharp

5. Joke Fight	kochelka@ufal.mff.cuni.cz	Peter, Aleš
---------------	---------------------------	-------------

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 doomscrolling 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.]

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)

The Classical Pipeline Strikes Back

(author)

- **The bright idea:** Can we generate useful game descriptions without LLMs? Traditional NLG uses a modular pipeline to turn structured data into text. Participants can build their own pipeline, from selecting relevant facts (content planning) and organizing them (document planning) to choosing wording (microplanning) and generating sentences (realization). For example, a description could highlight Linux compatibility and affordability for a *budget Linux gamer*, or VR support and demanding system requirements for a *hardcore VR enthusiast*. Participants can experiment with templates, rules, grammars, and different ways of selecting and presenting information to create descriptions that are both useful and personalized.
- **The goal:** Explore the fundamentals of classical NLG and generate game descriptions that are both readable (and personalized) while remaining faithful to the data.
- **Data to use:**
 - [Kaggle Steam Dataset 2025](#). Useful properties may include game name and description, price, release date, genres, tags, playtime, supported platforms, system requirements, and review sentiment. You don't need to use every property.
 - Participants can define their own player profiles to guide content selection.
- **Evaluation:** If the Tinder app is ready, the outputs could be evaluated there.

The Qualitative Vibe Check

(Kristýna, onderkova@ufal.mff.cuni.cz)

- **The bright idea:** Numbers tell us about a game, but not necessarily what it means. That comes from adding context for which we can use the LLMs' broad knowledge. Qualitative reasoning is a classic AI approach that abstracts numerical values and changes into meaningful qualitative states, such as *increasing*, *decreasing*, *stable*, or *approaching a critical threshold*. A \$70 game with overwhelmingly negative reviews might be a catastrophic AAA disappointment; a decade-old indie game with a huge player base might be a generational cult classic; a highly rated game with very few reviews might be a hidden gem; and something went wrong when a 2D pixel platformer demands an RTX 4090. Participants will build a system that identifies patterns in game data (combining numerical values, categories, and contextual knowledge) and turns them into entertaining insights.
- **The goal:** Explore whether combining qualitative reasoning with LLMs can produce more informative and nuanced insights rather than prompting an LLM to interpret raw numbers alone.

- **Data to use:**
 - [Kaggle Steam Dataset 2025](#). Useful properties include prices, release dates, genres, review sentiment and counts, platform compatibility, system requirements, and playtime information from user reviews. Not all properties need to be used.
 - Participants can combine properties, define their own qualitative categories and thresholds, and add contextual knowledge to generate interpretations.
- **Evaluation:** If the Tinder app is ready, the outputs could be evaluated there.

Who Wrote This? Stylometric Fingerprinting in Story Continuation

(Yongxin; yongxin.zhou@univ-grenoble-alpes.fr)

The bright idea: Give the same story prompt to humans and to several different LLMs, and each will continue it in its own way. Can you build a system that tells who wrote what — from style alone? Here's the twist: recent work shows model "fingerprints" are fragile. Pairing two texts works; single-text judgment collapses to near chance; change the prompt and accuracy drops. So the real challenge isn't *whether* fingerprints exist, it's *when* they become strong enough to be useful.

The goal: Build a system that attributes a story continuation to its source (human or which LLM), and figure out what makes attribution work or fail. You choose the features, the classifier, the evaluation, just make your choices explicit and comparable. If your system works, great. If it fails in an interesting way, that's also a result.

Data to use:

You have two ready-to-use datasets, each supporting a different question.

- **Path A:** Human vs AI attribution: [vkpriya/GPT-WritingPrompts](#) contains 97K prompts, each paired with human continuations (272K total) and GPT-3.5 continuations (206K total). No generation needed. Best for studying whether human style remains detectable.
- **Path B:** Multi-model attribution: [jirussell10/storyscope](#) contains 10K prompts with full stories from 5 LLMs (GPT-5.4, Claude Sonnet 4.6, DeepSeek V3.2, Kimi K2.5, Gemini 3 Flash) plus 304 pre-extracted narrative features. Note: this is **prompt-guided story generation** (write a story from scratch), not true continuation. Human text excluded for copyright. Best for comparing model-to-model fingerprints across more models.
- **Optional:** Full control over model choice: [euclaise/writingprompts](#) (or filtered [agentlans/euclaise-WritingPromptsX](#)) gives human continuations only; generate your own AI texts if you want models neither dataset covers.

Note: *StoryScope* already reports 68.4% macro-F1 for 6-way attribution using narrative features alone. Your task is to explore what it leaves open: surface stylometry alone, cross-topic

generalization, or human text as a sixth source.

Things you could explore (pick one or invent your own):

- **Features:** surface stylometry (sentence length, punctuation, TTR) vs POS n-grams vs dependency patterns vs perplexity/entropy
- **Conditions:** does attribution hold across prompt topics? Across text lengths (50/100/200/500 words)? Across paraphrase or translation?
- **Mixed authorship:** human opening + LLM ending (or vice versa), can you locate the switch point?
- **Classifier design:** logistic regression on handcrafted features vs fine-tuned BERT vs zero-shot prompting an LLM to guess
- **Evaluation:** accuracy, F1, but also — how do you handle class imbalance? What's your baseline?

Evaluation: Report Macro F1 for model attribution and human-vs-AI accuracy. Most importantly: **document your pipeline and share your outputs**, so we can compare across participants. A negative result (attribution near chance) is a valid finding if you can explain why.

Joke fight

(Peter Kocheika; kochelka@ufal.mff.cuni.cz)

- **The bright idea:** Create a system that when prompted will generate exactly five jokes about a given topic. Works like a tournament, in which pairs of Joke-generating teams fight.
- **The fight:** Both fighters propose a topic and have their system generate 5 jokes about both topics. Fighters then take turns reading jokes their system generated, and if the opponent laughs, they gain a point.
- **The goal:** Challenge your friends and have fun.
- **The survey:** <https://aclanthology.org/2025.inlg-main.45/>
- **The teaser:**

Why can't a T. rex clap?

Because it's extinct.