NORSK
REGNESENTRAL
NORWEGIAN
COMPUTING CENTER

Beyond the final text:

keystrokes as a window into language development and text production in the age of generative AI

Ildikó Pilán

3 September, 2026, Göteborg

Outline

- Introduction
- Keystroke data for **language proficiency profiling**
  - IncluEdS project overview
  - Automatic error correction and categorization
  - KEYSCORE study
- Keystroke data for **detecting AI-authored content**

Selina on Unsplash. Unsplash license.

NR in a nutshell

Norwegian Computing Center – a research institute for science and innovation

- Basic research projects
- Innovation projects
- Direct assignments (public and private sector)

Our research areas

→ Statistical modelling and machine learning

→ Statistical analysis of natural resource data

→ Applied research in information and communication technology

→ Image analysis and Earth observation

Webpage: https://nr.no/en/

NR in a nutshell

Norwegian Computing Center – a research institute for science and innovation

Our research areas

→ Statistical modelling and machine learning

→ Statistical analysis of natural resource data

→ Applied research in information and communication technology

→ Image analysis and Earth observation

- Basic research projects
- Innovation projects
- Direct assignments (public and private sector)

NLP "group"

Pierre Lison

Ildikó Pilán

Mark Anderson

Max Hort

Webpage: https://nr.no/en/

Motivation

- Generative AI **transforms** how we write and **challenges** traditional notions of **authorship** and **assessment**.

UK universities warned to 'stress-test' assessments as 92% of students use AI

Survey of 1,000 students shows 'explosive increase' in use of generative AI in particular over past 12 months

Link: https://www.theguardian.com/education/2025/feb/26/uk-universities-warned-to-stress-test-assessments-as-92-of-students-use-ai?utm_source=chatgpt.com

Motivation

- Generative AI **transforms** how we write and **challenges** traditional notions of **authorship** and **assessment**.

- There is an increasing interest in methods and tools that look **beyond the final text** and learn from the **writing process** itself.
    - Detailed information about **how** the text was produced
    - Indicators of cognitive **effort** (or the lack thereof), **hesitation** etc.

Keystroke log (KL) data


```
"1761825273889": "keydown: t",
"1761825273986": "keydown: h",
"1761825273993": "keyup: t",
"1761825274033": "keydown: i",
"1761825274066": "keyup: h",
"1761825274113": "keyup: i",
"1761825274201": "keydown: s",
"1761825274257": "keydown:  ",
"1761825274266": "keyup: s",
"1761825274313": "keyup:  ",
"1761825274457": "keydown: t",
"1761825274498": "keydown: i",
"1761825274546": "keyup: t",
"1761825274586": "keyup: i",
"1761825274746": "keydown: l",
"1761825274796": "keyup: l",
"1761825274881": "keydown: l",
"1761825274946": "keyup: l",
"1761825275146": "keydown: Backspace",
"1761825275194": "keyup: Backspace",
"1761825275265": "keydown: Backspace",
"1761825275334": "keyup: Backspace",
"1761825275425": "keydown: Backspace",
"1761825275458": "keyup: Backspace",
```


Unix timestamp

Raw logs can be “summarized” (linearized)

This is a <0.576>test <1.184>text for <0.696>this till<DEL3> ool. <3.680>I’m su< DEL2 > <0.961>just making up some text.

These show:

Pauses
Revisions
Burst

Bursts: Continuous stretches of writing between pauses and revisions.

Examples of time-based indicators from keystrokes

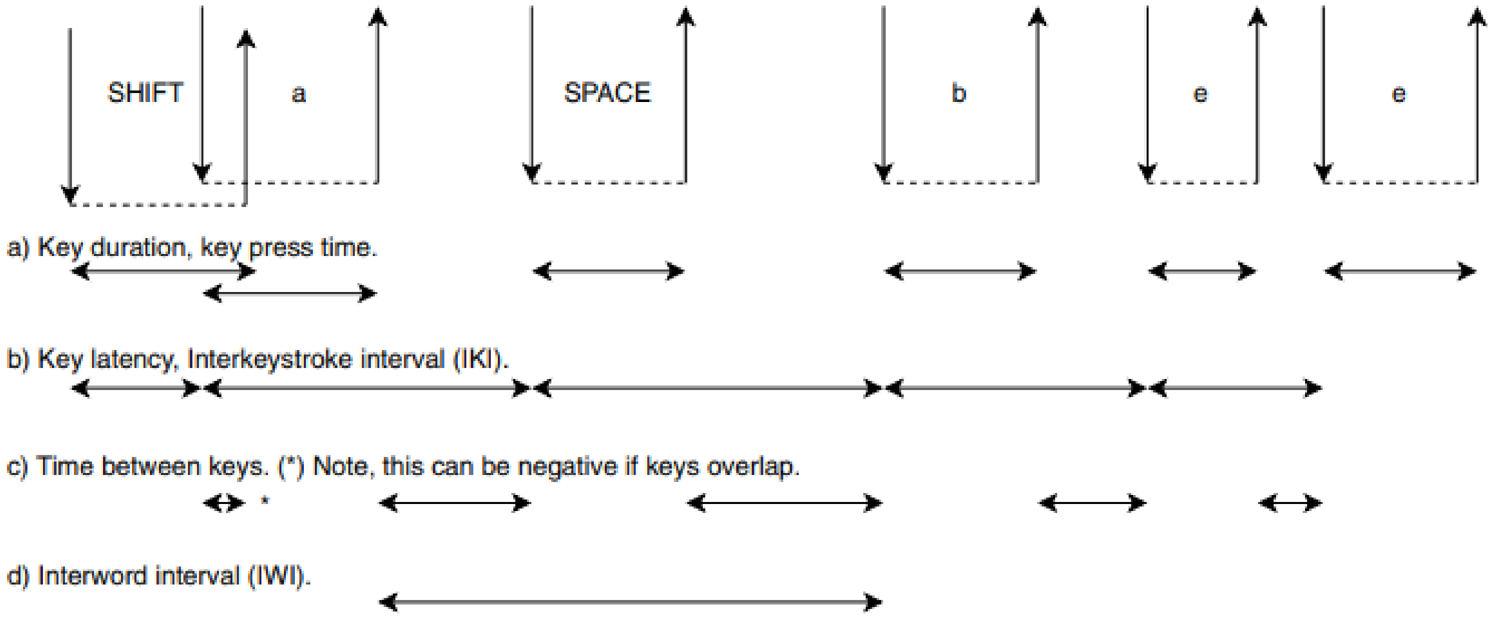


Figure from: https://link.springer.com/article/10.1007/s11145-019-09953-8

Challenges for wider adoption

- **Privacy**:
  - keystroke logs are more revealing (includes deleted content)
  - possibly identifying (used also for login verification)
- **Technical** and **methodological**:
  - distinguish between cognitive processes and noise at scale,
  - adapt to variability in writers, (educational) settings, devices
- **Trust** and **usability**:
  - reliable, actionable insights for a variety of (pedagogical) purposes

Keystroke data for language proficiency profiling

The IncluEdS project

- **IncluEdS:** Keystroke-Based Language Proficiency Profiling for **Inclu**sive **Ed**ucation and **S**ociety (2026 –2029, Norwegian Research Council)

- Partners:

NORSK REGNESENTRAL NORWEGIAN COMPUTING CENTER

Norwegian Reading Centre
National Centre for Reading Education and Research
University of Stavanger

UNIVERSITETET I OSLO

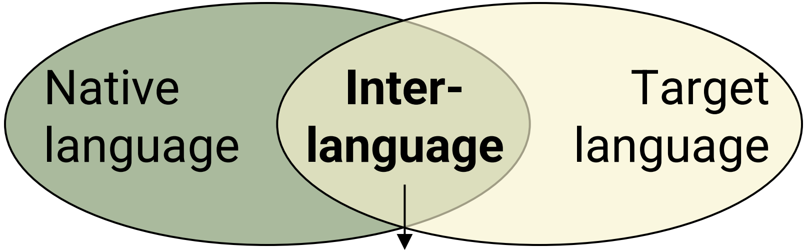
UNIVERSITY OF GOTHENBURG

- P(ilán) I(ldikó) ☺

IncluEdS: objectives

- **Collect** a **keystroke-log dataset** written by learners of **Norwegian** as a second language (L2).
- Compare learners **pausing** and **revision** behavior at **different** L2 **levels**.
- Leverage this data in interpretable **machine learning models** for **L2 writing assessment**.
- **Test** the **reliability** & pedagogical **usefulness** of these methods.

Second Language Acquisition (SLA) and CEFR



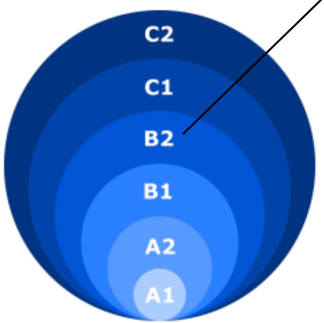
Shows systematic **developmental patterns**

- *Subject-Verb agreement*
- *Noun phrase agreement*
- *Plural*

Order of acquisition of linguistic structures. (Pienemann, 2005).

EU: **CEFR**

(*Common European Framework of Reference for Languages*)



B2: *Has a **good command** of simple language structures and some complex grammatical forms, although they tend to use complex structures rigidly with some inaccuracy.**

*Descriptor of grammatical accuracy (linguistic competence)

Keystroke log (KL) data


```
"1761825273889": "keydown: t",
"1761825273986": "keydown: h",
"1761825273993": "keyup: t",
"1761825274033": "keydown: i",
"1761825274066": "keyup: h",
"1761825274113": "keyup: i",
"1761825274201": "keydown: s",
"1761825274257": "keydown:  ",
"1761825274266": "keyup: s",
"1761825274313": "keyup:  ",
"1761825274457": "keydown: t",
"1761825274498": "keydown: i",
"1761825274546": "keyup: t",
"1761825274586": "keyup: i",
"1761825274746": "keydown: l",
"1761825274796": "keyup: l",
"1761825274881": "keydown: l",
"1761825274946": "keyup: l",
"1761825275146": "keydown: Backspace",
"1761825275194": "keyup: Backspace",
"1761825275265": "keydown: Backspace",
"1761825275334": "keyup: Backspace",
"1761825275425": "keydown: Backspace",
"1761825275458": "keyup: Backspace",
```


Unix timestamp

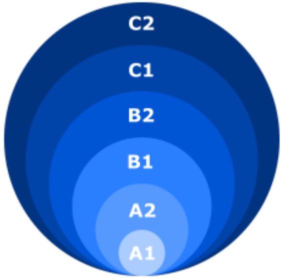
Raw logs can be "summarized" (linearized)

This is a <0.576>test <1.184>text for <0.696>this till<DEL3> ool. <3.680>I'm su< DEL2> <0.961>just making up some text.

Pauses
Revisions
Burst

+ **NLP tools** →

CEFR



Bursts: Continuous stretches of writing between pauses and revisions.

IncluEdS: ongoing data collection

- **Participants:**
  - L2 Norwegian learners
  - 16+ years old,
  - A2-B2 levels mainly
- **Tasks** (ca. 90 min. in total):
  - Background survey
  - Copy task
  - Essay task (45 min)
- **Setting**: in classroom, no external resources are allowed

ScriptLog: copy task

OPPTAK | AVSPILLING | ANALYSE | INNSTILLINGER

KODE: **111111**

OPPGAVE-ID: **copy**

START | STOPP

Ord: 0

Kopieringsoppgave:

Skriv den første setningen fort 7 ganger, én per linje:
- gratulerer med dagen kjære onkel

Skriv den andre setningen fort 7 ganger, én per linje:
- hyggelig å møte deg

Ikke rett små feil. Ikke bruk store bokstaver eller punktum.

Klikk START.

Skriv deretter teksten din her.

Klikk STOPP når du er ferdig.

gratulerer med dagen kjære onkel
gratulerer med dagen kjære onkle
gratulerer med dagen kjære onkel
gratulerer med daen kjære onlkel
grstulerer med dsgen kjære onkel
gratulerer med dagen kjære onkel
gratulere med dagen kjære onkel
hyggelig å møte deg
hyggeleig å møte deg
hyggelig å møte deg
hyggelig å møte deg
hyggelig å møte sdeg
hyggelig å møte deg
hyggelig å møte deg

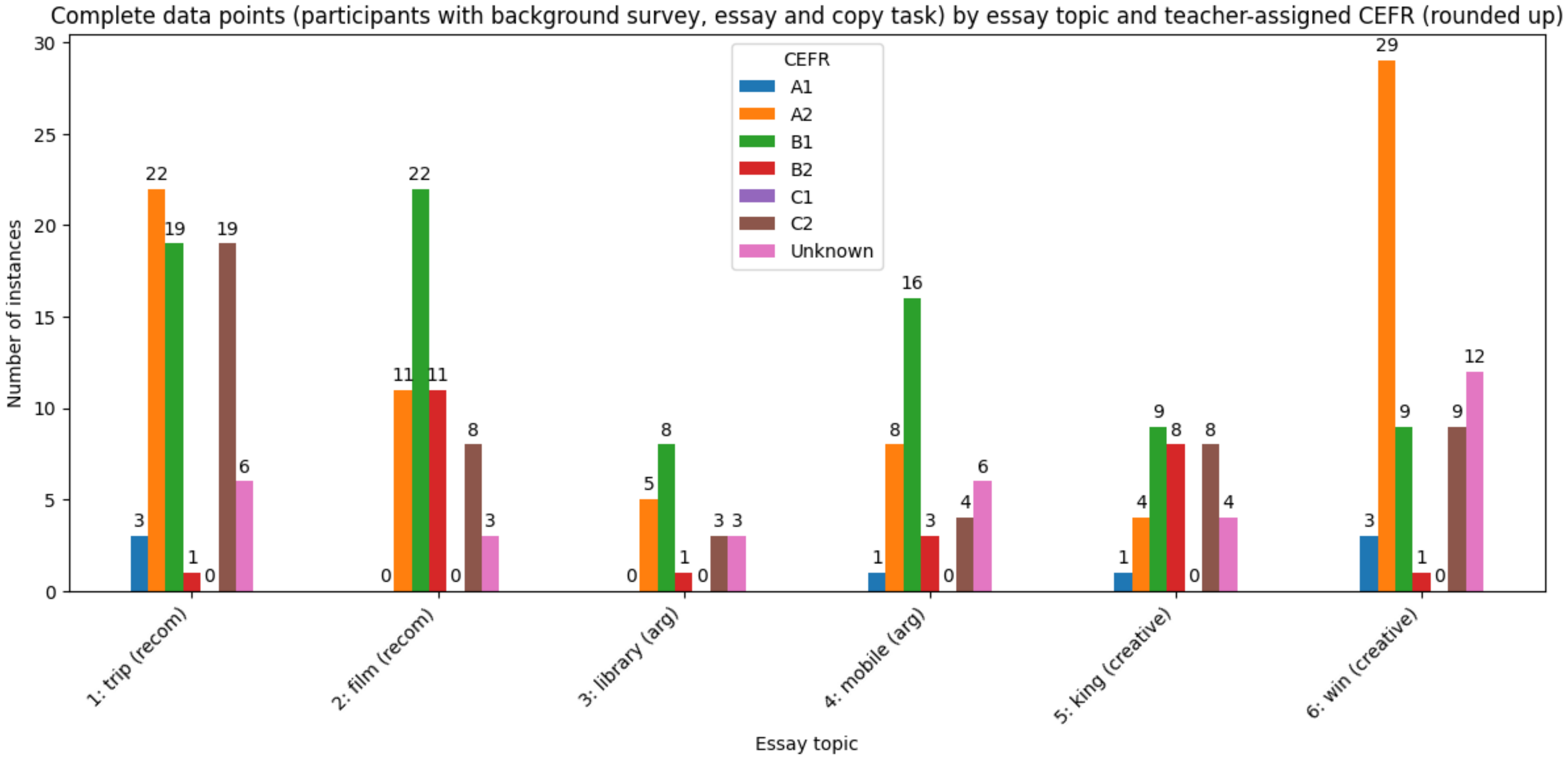
ScriptLog was adapted to the project by Johan Frid, Humanities Lab, University of Lund.

ScriptLog: essay task

Essay topics
1. Recommendation: excursion
2. Recommendation: film
3. Argumentative: library closing
4. Argumentative: no mobiles in school
5. Creative: king/queen for a day
6. Creative: winning money

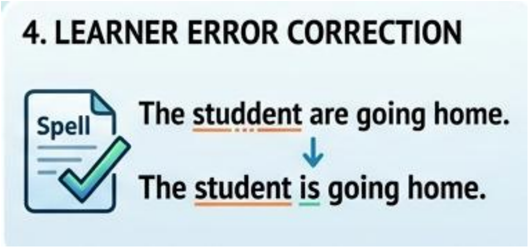
ScriptLog was adapted to the project by Johan Frid, Humanities Lab, University of Lund.

About 300 essays collected so far



Error correction and categorization

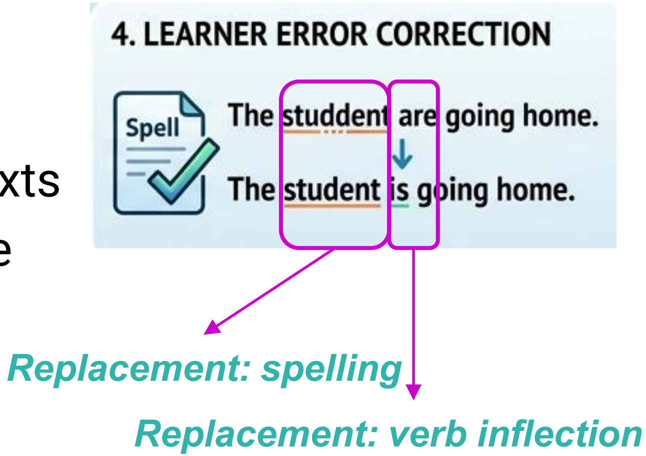
- LLM-based automatic **error correction**
  - Based on 5 example sentences
  - Evaluated on 12 manually corrected IncluEdS texts
  - <u>To do</u>: experiment with more LLMs + re-evaluate



Work carried out in large part by a summer intern at NR, Peder Mæhle Førland.

Error correction and categorization

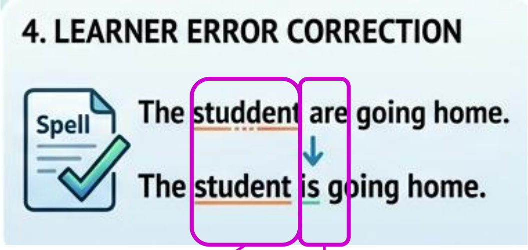
- LLM-based automatic **error correction**
  - Based on 5 example sentences
  - Evaluated on 12 manually corrected IncluEdS texts
  - To do: experiment with more LLMs + re-evaluate
- Rule-based **edit extraction** + **error categorization**
  - Adapted the ERRANT tool to Norwegian



Work carried out in large part by a summer intern at NR, Peder Mæhle Førland.

Error correction and categorization

- LLM-based automatic **error correction**
  - Based on 5 example sentences
  - Evaluated on 12 manually corrected IncluEdS texts
  - <u>To do</u>: experiment with more LLMs + re-evaluate
- Rule-based **edit extraction** + **error categorization**
  - Adapted the [ERRANT]() tool to Norwegian
  - Evaluation on the **ASK corpus** (L2 Norwegian)
    - Mapping of ERRANT to ASK error categories
  - <u>To do</u>: evaluate on IncluEdS essays



Replacement: spelling

Replacement: verb inflection

ORT *(Orthography)*

INFL *(Wrong inflected form)*

Work carried out in large part by a summer intern at NR, Peder Mæhle Førland.

How well do these automatic approaches work?

- **Error categorization**:
  - Quite well in general
  - Differences between categories

Category	Precision	Recall	F0.5
Correct span	0.919	0.893	0.913
Correct span + tag (All tags)	0.796	0.774	0.791

Work carried out in large part by a summer intern at NR, Peder Mæhle Førland.

How well do these automatic approaches work?

- **Error categorization**:
  - Quite well in general
  - Differences between categories

Category	Precision	Recall	F0.5
Correct span	0.919	0.893	0.913
Correct span + tag (All tags)	0.796	0.774	0.791

- **Correction**:
  - some under- and overcorrection -> we will try different prompts & LLMs

```
SAMPLE 5
SOURCE: De ikke snakke, de er ofte elene vi må hjeipe hverandre.
PRED:   De kan ikke snakke, de er ofte alene og vi må hjelpe hverandre.
REF:    De snakker ikke, de er ofte alene og vi må hjelpe hverandre.
```

```
SAMPLE 6
SOURCE: Jeg tror på turen den jeg likte best å spice og å gå.
PRED:   Jeg tror på turen var det jeg likte best å spise og å gå.
REF:    Jeg tror det jeg likte best på turen var å spise og å gå.
```

Examples of automatic (PRED) and manual (REF) corrections

Work carried out in large part by a summer intern at NR, Peder Mæhle Førland.

Upcoming plans

- Manual annotation: **CEFR-levels**
  - Planned both based on the **final** texts and based on **process-enriched** version (with pause and revision information)
  - Inclusion of short **explanations**
- Manual annotation: **revision** type or intention
- Computational modeling: **explainable CEFR-level prediction**
- Quantitative SLA study: e.g. order of acquisition

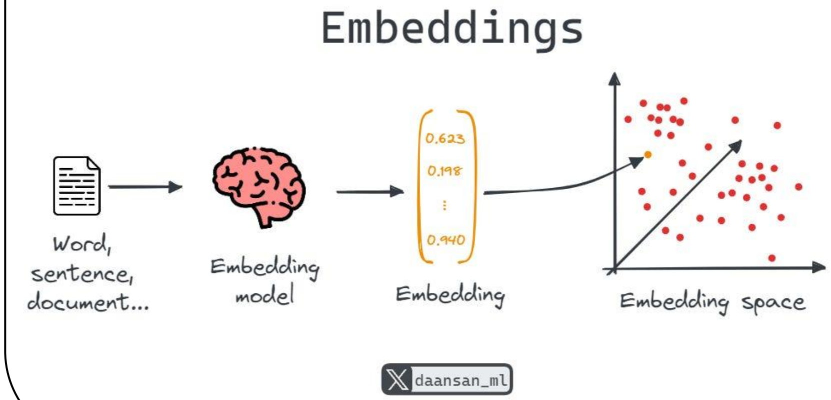
KEYSCORE - Keystroke-enhanced Automated Essay Scoring (Schaller et al., 2026)

- RQ: **How early** in the writing process can we give meaningful feedback (=score) to learners?
- Task: **Essay scoring** (L2 English) with a **transformer** model (DeBERTa)
- Tested on **intermediate snapshots** at 5, 10, 15, 20 and 25 minutes of writing
- 2 conditions: essays written **with** and **without** the use of **AI chatbots**
- Tested **text features** and **process features** separately and together
  - Text f: vector representation of the text
  - Process f: mean pause duration, deletion / total keys etc.

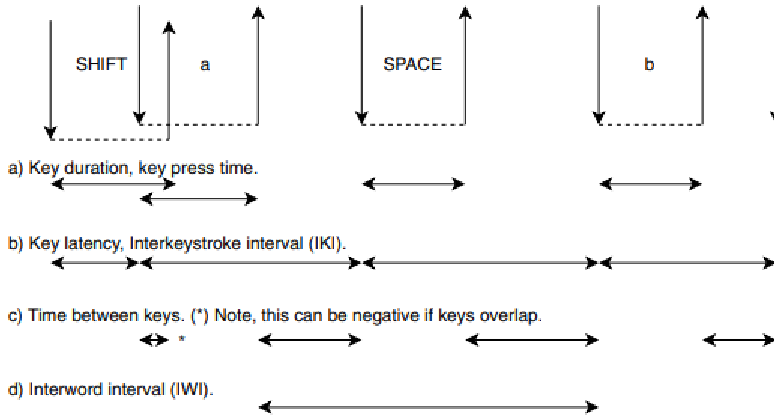
Link: https://aclanthology.org/2026.bea-1.24/

Combined feature set: text embeddings + keystrokes

Neural text representation

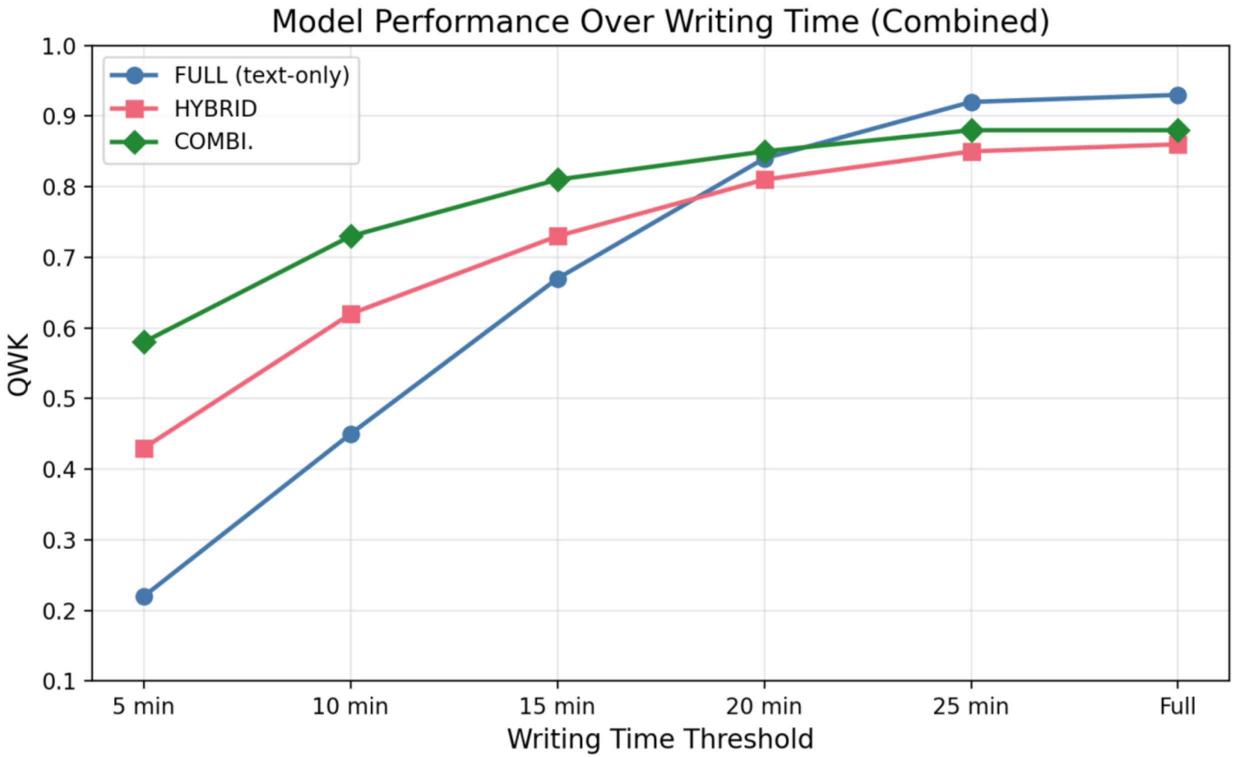


Keystroke features



[0.42, 0.17, 0.83, 0.31, 0.65, …] + *[0.73, 0.21, 0.56, 0.89, …]*

KEYSCORE: results



- Process-features useful for scoring **early** (especially when chatbot used)

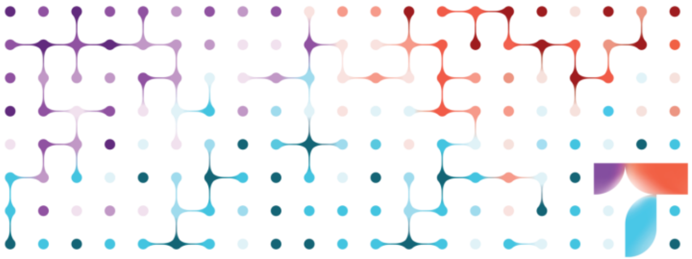
Automatic detection of AI-authored content

Connected to two research projects

TRUST AI Center

The Norwegian Centre for Trustworthy AI

Building the foundations for AI that is accurate, interpretable, inclusive, fair, safe, sustainable, and well-governed.



DetectAI (TRUST sub-project)
PI: Pierre Lison
2026-2030

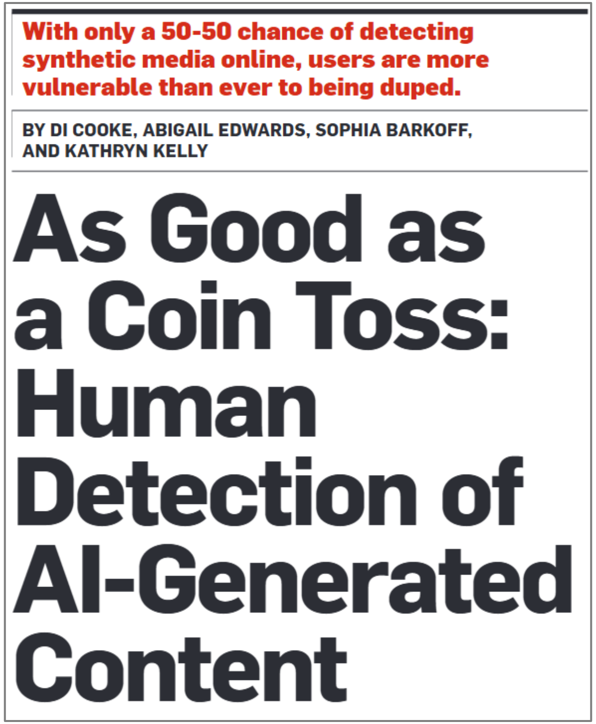
Artificial Intelligence: Human-LLM Communication (**AICOM**) project (UiO)

Ingrid Lossius Falkum
University of Oslo
Principal investigator

Nicholas Allott
University of Oslo
Co-Principal investigator

Detecting AI-use is hard

- **Human performance** remains **close to chance level**, particularly if AI-content is **paraphrased.** (Mehta et al., 2026)
- **Automatic methods** have several **shortcomings.** (Fraser et al., 2025)
- **Behavioral** data helps (Mehta et al., 2026)
  - But **threat models** can mimic natural keystroke dynamics…
  - Model resilience can be improved by **adversarial training** (training also on such deceptive data).


With only a 50-50 chance of detecting synthetic media online, users are more vulnerable than ever to being duped.

BY DI COOKE, ABIGAIL EDWARDS, SOPHIA BARKOFF, AND KATHRYN KELLY

As Good as a Coin Toss: Human Detection of AI-Generated Content

https://dl.acm.org/doi/epdf/10.1145/3729417

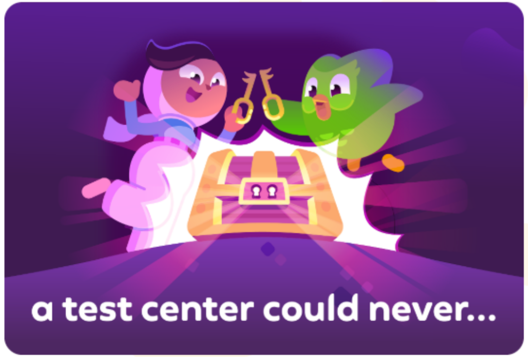
Current work: scoring suspicion of AI Use

- Experimented with heuristics for flagging **suspicious text segments** in keystroke logs based on e.g.:
  - Paste-like indicators
    - large **text jump** in a very **short time**
    - very **few keydown** events for a large insertion
  - Copy-typing (or automated typing) indicators
    - **many characters** typed **quickly**
    - **«monotonous» writing**
      - low variance in inter-key intervals
      - almost **no corrections**
- Suspiciousness is estimated using rule-based thresholds.
- Assigns a **suspicion score** so candidate segments can be **reviewed** manually.

APRIL 10, 2025

Why can't test centers catch cheating effectively?

Sophie Wodzak



How the DET catches what test centers miss

Unlike traditional exams, the DET is fully recorded and analyzed using advanced technology to flag even the most discreet forms of misconduct. Here's how:

AI-powered keystroke monitoring detects unnatural typing

One of the biggest security challenges in language testing is detecting when a test taker isn't using their own words. The DET's AI tracks how responses are typed to flag suspicious behavior, such as:

- Copy-typing from an external source—If someone is transcribing text rather than thinking and writing naturally, AI can detect unusual typing rhythms and speed changes.
- Abrupt shifts in writing style—If part of an essay is written fluently but another part has unusual errors, it could indicate partial AI assistance or memorized content.
- Pausing in unnatural places—Frequent pauses at odd moments could suggest that a test taker is referring to external notes.

Unlike human proctors, who only see the final written response, the DET tracks how that response was written—making it much harder to cheat undetected.

https://blog.englishtest.duolingo.com/why-cant-test-centers-catch-cheating-effectively/

Rethinking writing assessment for the AI era

- Was AI used? ⇨ **How** was AI used?
- Develop assessment methods that capture **collaborative human-AI writing processes.**
- Support and identify **constructive** practices in AI-assisted writing and its computational analysis via keystroke logs.

Thank you!

(Some) References

- *Cooke, D., Edwards, A., Barkoff, S., & Kelly, K. (2025). As good as a coin toss: human detection of ai-generated content. Communications of the ACM, 68(10), 100-109.*
- *Fraser, K. C., Dawkins, H., & Kiritchenko, S. (2025). Detecting ai-generated text: Factors influencing detectability with current methods. Journal of Artificial Intelligence Research, 82, 2233-2278.*
- *Mehta, A., Kumar, R., Singla, A., Bisht, K., Singla, Y. K., & Shah, R. R. (2026). Detecting LLM-assisted academic dishonesty using keystroke dynamics. IEEE Transactions on Biometrics, Behavior, and Identity Science.*
- *Pienemann, M., Di Biase, B., & Kawaguchi, S. (2005). Extending processability theory. Cross-linguistic aspects of Processability Theory, 30, 199-251.*
- *Schaller, N. J., Mora, D., Jansen, T., Köller, O., & Horbach, A. (2026) KEYSCORE—Keystroke-enhanced Automated Essay Scoring. Proceedings of BEA 2026. ACL.*
- *Velentzas, G., Caines, A., Borgo, R., Pacquetet, E., Hamilton, C., Arnold, T., ... & Yannakoudakis, H. (2024). Logging keystrokes in writing by English learners. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 10725-10746).*