– IN5550 –
Neural Methods in Natural Language Processing
Home Exam: Task Overview and Kick-Off

Jindřich Helcl, Andrey Kutuzov, Vladislav Mikhailov, Meihan Tong,

Shenbin Qian, David Samuel, Yves Scherrer

University of Oslo

April 27th, 2026



Contents

1. Introduction
2. 1: English – Sámi machine translation
3. 2: Disentangling content and formal reasoning in LLMs
4. 3: Predicting Emotional Valence and Arousal over Time
5. 4: Open-Ended Regional Question Answering with RAG
6. 5: Steering LLMs responses towards moral foundations
7. 6: Machine-generated Code Detection
8. What's next?

You made it!

Congratulations to those qualified for the exam!

- IN9550: 5 students qualified (out of 8)
- IN5550: 39 students qualified (out of 56)

You made it!

Congratulations to those qualified for the exam!

- IN9550: 5 students qualified (out of 8)
- IN5550: 39 students qualified (out of 56)

About 63-69 % 'qualification rate'.

You made it!

Congratulations to those qualified for the exam!

- IN9550: 5 students qualified (out of 8)
- IN5550: 39 students qualified (out of 56)

About 63-69 % 'qualification rate'.
Now it's time to take the exam!

All home exam info is on the semester page

Between Monday April 27 and Monday June 1, there will be a home exam in the form of a small research project. To pass the exam, students need to submit and have accepted a scientific paper to the 8th IN5550 Workshop on Neural Natural Language Processing (WNNLP 2026).

Background

Final examination in this class takes the form of a 'home exam', i.e. a project that students can work on over a period of three weeks. Like for the exam-qualifying obligatory assignments, group work (in teams of up to three students) is encouraged (but **PhD-students** enrolled for **IN9550** must complete the home exam **individually**). Team composition needs to be declared at the start of the exam period and cannot be changed after April 29th, 2026. Also, each team needs to decide beforehand which of the available topics they want to research; these 'tracks' of the exam will be introduced in the lecture on April 27. Further background on the tracks and supporting data and code will be available through the course GitHub repo. Please announce your team composition and choice of track no later than April 29th by emailing the course contact address `in5550-help@ifi.uio.no`.

The main exam period will be Monday, April 27 to Wednesday, May 20, 2026. Once a team declares their track (by email, see above) they can start working on it, using a written set of suggestions for the specific track available through the GitHub repository. During the exam period, there will be no regular lectures or group sessions. However, project teams will be offered mentoring sessions with the track chairs (the course instructors); these sessions will be organized as a joint meeting for all teams subscribed to a given track.

https://www.uio.no/studier/emner/matnat/ifi/IN5550/v26/final-exam/

8th IN5550 Workshop on Neural NLP (WNNLP 2026)

General Idea

- Preparing a scientific paper for publication.
- NB: not another obligatory assignment! You have to research things.

8th IN5550 Workshop on Neural NLP (WNNLP 2026)

General Idea

- Preparing a scientific paper for publication.
- NB: not another obligatory assignment! You have to research things.

Standard life cycle of a scientific contribution:

(0) Problem Statement

(1) Experimentation

(2) Analysis

(3) Paper Submission

(4) Reviewing

(5) Submission of the Final ('Camera-Ready') Version

(6) Presentation

WNNLP 2026: Call for Papers

Most important Constraints

- Six specialized tracks with their track chairs:
  1. English – Sámi machine translation (Yves Scherrer and David Samuel)
  2. Disentangling content and formal reasoning in LLMs (Meihan Tong)
  3. Predicting Emotional Valence and Arousal over Time (Shenbin Qian)
  4. Open-Ended Regional Question Answering with RAG (Jindřich Helcl)
  5. Steering LLMs Responses Towards Moral Foundations (Hannah Clausen and Andrey Kutuzov, original proposal by Christian Grusd)
  6. Machine-generated code detection (Vladislav Mikhailov)

WNNLP 2026: Call for Papers

Most important Constraints

- Six specialized tracks with their track chairs:
  1. English – Sámi machine translation (Yves Scherrer and David Samuel)
  2. Disentangling content and formal reasoning in LLMs (Meihan Tong)
  3. Predicting Emotional Valence and Arousal over Time (Shenbin Qian)
  4. Open-Ended Regional Question Answering with RAG (Jindřich Helcl)
  5. Steering LLMs Responses Towards Moral Foundations (Hannah Clausen and Andrey Kutuzov, original proposal by Christian Grusd)
  6. Machine-generated code detection (Vladislav Mikhailov)
- You choose your track (max 6 teams per track, first come - first served)
- Today we present the tracks only briefly; check the course GitHub repo for detailed descriptions, datasets and pre-code

WNNLP 2026: Call for Papers

Most important Constraints

- Six specialized tracks with their track chairs:
  1. English – Sámi machine translation (Yves Scherrer and David Samuel)
  2. Disentangling content and formal reasoning in LLMs (Meihan Tong)
  3. Predicting Emotional Valence and Arousal over Time (Shenbin Qian)
  4. Open-Ended Regional Question Answering with RAG (Jindřich Helcl)
  5. Steering LLMs Responses Towards Moral Foundations (Hannah Clausen and Andrey Kutuzov, original proposal by Christian Grusd)
  6. Machine-generated code detection (Vladislav Mikhailov)
- You choose your track (max 6 teams per track, first come - first served)
- Today we present the tracks only briefly; check the course GitHub repo for detailed descriptions, datasets and pre-code
- Long papers: up to 8 pages (minimally 5), excluding references, in ACL Rolling Review LaTeX template.
- Submitted papers must be anonymous:
  - peer reviewing is double-blind.

WNNLP 2026: Important Dates

Schedule

By April 29	Declare choice of track (and team composition)
Week 19 or 20	Track specific joint mentoring meetings
May 20	(Strict) Submission deadline for your papers
May 21–25	Reviewing period: Each student reviews two papers
May 27	Track Chairs announce acceptance decisions
May 28	Camera-ready papers due, with requested revisions
June 1	Oral presentations and awards at the workshop

- All dates AOE (anywhere on Earth).
- Submission website:
  - `https://openreview.net/group?id=uio.no/Univeristy_Of_Oslo/2026/WNNLP`
- Replicability: submission must be backed by code repository (visible to track chairs only).

WNNLP 2026: Programme Committee

General Chair

- Andrey Kutuzov

Track Chairs

- English – Sámi machine translation: Yves Scherrer and David Samuel
- Disentangling content and formal reasoning in LLMs: Meihan Tong
- Predicting Emotional Valence and Arousal over Time: Shenbin Qian
- Open-Ended Regional Question Answering with RAG: Jindřich Helcl
- Steering LLMs Responses Towards Moral Foundations: Hannah Clausen and Andrey Kutuzov
- Machine-generated Code Detection: Vladislav Mikhailov

Peer Reviewers

- All students who have submitted a scientific paper (you)

Contents

1. Introduction
2. 1: English – Sámi machine translation
3. 2: Disentangling content and formal reasoning in LLMs
4. 3: Predicting Emotional Valence and Arousal over Time
5. 4: Open-Ended Regional Question Answering with RAG
6. 5: Steering LLMs responses towards moral foundations
7. 6: Machine-generated Code Detection
8. What's next?

1: English – Sámi machine translation

Main info

Track chairs: Yves Scherrer and David Samuel

Learn more:

`https://github.uio.no/in5550/2026/tree/main/exam/saami_mt`

Background

WMT is the yearly Conference on Machine Translation and is known for their shared tasks on translation.

Background

WMT is the yearly Conference on Machine Translation and is known for their shared tasks on translation.

- Each year, WMT shared tasks focus on new language pairs
- In 2026 (and for the first time in its history), an official language of Norway is included: English $\rightarrow$ Northern Sámi
- `https://www2.statmt.org/wmt26/translation-task.html`

Background

WMT is the yearly Conference on Machine Translation and is known for their shared tasks on translation.

- Each year, WMT shared tasks focus on new language pairs
- In 2026 (and for the first time in its history), an official language of Norway is included: English $\rightarrow$ Northern Sámi
- `https://www2.statmt.org/wmt26/translation-task.html`

Your task: develop a machine translation system that translates **from English to Northern Sámi**, following the guidelines and requirements of WMT.

Background

WMT is the yearly Conference on Machine Translation and is known for their shared tasks on translation.

- Each year, WMT shared tasks focus on new language pairs
- In 2026 (and for the first time in its history), an official language of Norway is included: English → Northern Sámi
- `https://www2.statmt.org/wmt26/translation-task.html`

Your task: develop a machine translation system that translates **from English to Northern Sámi**, following the guidelines and requirements of WMT.

- Knowledge of (Northern) Sámi is an advantage, but by no means a requirement!

The challenges

Northern Sámi is a very low-resource language.

- There are very few available datasets for training, and they are small and noisy.
- You may need to spend considerable efforts on **curating and augmenting training data**.

The challenges

Northern Sámi is a very low-resource language.

- There are very few available datasets for training, and they are small and noisy.
- You may need to spend considerable efforts on **curating and augmenting training data**.

Most pre-trained models don't 'know' Northern Sámi.

- But new language pairs can be added by fine-tuning.
- You will experiment with different base models and adaptation strategies.
- You can use task-specific encoder-decoder models or general-purpose decoder-only LLMs.

The challenges

Northern Sámi is a very low-resource language.

- There are very few available datasets for training, and they are small and noisy.
- You may need to spend considerable efforts on **curating and augmenting training data**.

Most pre-trained models don't 'know' Northern Sámi.

- But new language pairs can be added by fine-tuning.
- You will experiment with different base models and adaptation strategies.
- You can use task-specific encoder-decoder models or general-purpose decoder-only LLMs.

WMT has chosen the harder translation direction.

- Translating from Northern Sámi to English is presumably easier, but this is not the goal of the task.

Contents

1. Introduction
2. 1: English – Sámi machine translation
3. 2: Disentangling content and formal reasoning in LLMs
4. 3: Predicting Emotional Valence and Arousal over Time
5. 4: Open-Ended Regional Question Answering with RAG
6. 5: Steering LLMs responses towards moral foundations
7. 6: Machine-generated Code Detection
8. What's next?

2: Disentangling content and formal reasoning in LLMs

Main info

Track chairs: Meihan Tong

Learn more: `https://github.uio.no/in5550/2026/tree/main/exam/disentangling`

Track 2: Task Overview

Background

- LLM performance often reflects a mixture of content knowledge and formal reasoning.
- Many benchmarks do not clearly separate the two.

Track 2: Task Overview

Background

- LLM performance often reflects a mixture of content knowledge and formal reasoning.
- Many benchmarks do not clearly separate the two.

Task

- Design experiments to disentangle reasoning from content in LLMs.
- Study how models behave when:
  - reasoning structure is fixed but content changes;
  - content is similar but reasoning demands differ.

Track 2: Task Overview

Background

- LLM performance often reflects a mixture of content knowledge and formal reasoning.
- Many benchmarks do not clearly separate the two.

Task

- Design experiments to disentangle reasoning from content in LLMs.
- Study how models behave when:
  - reasoning structure is fixed but content changes;
  - content is similar but reasoning demands differ.

Evaluation

- Measure accuracy, content effect, and combined score.
- Analyze whether errors come from weak reasoning, missing knowledge, or both.

Contents

1. Introduction
2. 1: English – Sámi machine translation
3. 2: Disentangling content and formal reasoning in LLMs
4. 3: Predicting Emotional Valence and Arousal over Time
5. 4: Open-Ended Regional Question Answering with RAG
6. 5: Steering LLMs responses towards moral foundations
7. 6: Machine-generated Code Detection
8. What's next?

3: Predicting Emotional Valence and Arousal over Time

Main info

Track chairs: Shenbin Qian

Learn more:

`https://github.uio.no/in5550/2026/tree/main/exam/valence`

Track 3: Task Overview

Goal

Given a short personal text, predict two emotional scores:

- Valence: how positive/negative (-2 to $+2$)
- Arousal: how energized/calm (0 to 2)

Track 3: Task Overview

Goal

Given a short personal text, predict two emotional scores:

- Valence: how positive/negative (-2 to $+2$)
- Arousal: how energized/calm (0 to 2)

Data: Ecological Momentary Assessment (EMA)

- Participants wrote texts over time describing their emotional state
- Two text types: **essays** ($\sim$59 words) and **feeling-word lists** ($\sim$5 words)
- Train: 2,764 texts from 137 authors; Test: 1,737 texts from 91 authors
- Longitudinal: multiple entries per user across up to 7 collection phases

Track 3: Task Overview

Goal

Given a short personal text, predict two emotional scores:

- Valence: how positive/negative (-2 to $+2$)
- Arousal: how energized/calm (0 to 2)

Data: Ecological Momentary Assessment (EMA)

- Participants wrote texts over time describing their emotional state
- Two text types: **essays** ($\sim$59 words) and **feeling-word lists** ($\sim$5 words)
- Train: 2,764 texts from 137 authors; Test: 1,737 texts from 91 authors
- Longitudinal: multiple entries per user across up to 7 collection phases

What makes this different from regular sentiment analysis?

- Two independent dimensions, not just positive/negative
- Evaluation rewards tracking within-person variation over time

Track 3: Evaluation

Three correlation metrics (computed for valence and arousal separately)

- Within-person (r_{within}): average of per-user Pearson r
  - How well do you track one individual's emotional fluctuations?
- Between-person (r_{between}): Pearson r across user-level means
  - Can you distinguish generally happier users from sadder ones?
- Composite ($r_{\text{composite}}$): Fisher's z-transform combination
  - $r_{\text{composite}} = \tanh\left(\frac{\text{arctanh}(r_{\text{within}}) + \text{arctanh}(r_{\text{between}})}{2}\right)$
  - **Primary ranking metric**

Origin

Based on SemEval-2026 Task 2 (Subtask 1). Dataset and evaluation scripts provided.

Contents

1. Introduction
2. 1: English – Sámi machine translation
3. 2: Disentangling content and formal reasoning in LLMs
4. 3: Predicting Emotional Valence and Arousal over Time
5. 4: Open-Ended Regional Question Answering with RAG
6. 5: Steering LLMs responses towards moral foundations
7. 6: Machine-generated Code Detection
8. What's next?

4: Open-Ended Regional Question Answering with RAG

Main info

Track chair: Jindřich Helcl

Learn more: `https://github.uio.no/in5550/2026/tree/main/exam/regional-qa`

4: Open-Ended Regional Question Answering with RAG

Background

- Most QA benchmarks focus on globally known facts and use multiple-choice formats.
- Regional knowledge is a particularly hard test case: culturally specific facts from less-represented languages, often absent from model training data.
- **CUS-QA**: open-ended QA benchmark in Czech, Slovak, and Ukrainian, grounded in local Wikipedias. Even the best open-weight LLMs achieve only $\approx 60\%$ accuracy.

4: Open-Ended Regional Question Answering with RAG

Task

- Build a **cross-lingual RAG system** that retrieves relevant Wikipedia passages and generates correct open-ended answers in all three languages.

Required experiments

- **Oracle RAG:** use the reference Wikipedia URL as context — upper bound on retrieval quality.
- **Embedding-based RAG:** build a dense index over FineWiki; baseline embedding model is Multilingual-E5-Large.
- Compare multiple open-weight LLMs as the generator.

Evaluation

- Primary: LLM-as-a-judge (M-Prometheus) on the dev set.
- Test set scored on Codabench (chrF, BLEU, BERTScore, . . .).

Contents

1. Introduction
2. 1: English – Sámi machine translation
3. 2: Disentangling content and formal reasoning in LLMs
4. 3: Predicting Emotional Valence and Arousal over Time
5. 4: Open-Ended Regional Question Answering with RAG
6. 5: Steering LLMs responses towards moral foundations
7. 6: Machine-generated Code Detection
8. What's next?

5: Steering LLMs responses towards moral foundations

Main info

Track chairs: Hannah Clausen and Andrey Kutuzov

Learn more:

`https://github.uio.no/in5550/2026/tree/main/exam/MFQ`

5: Steering LLMs responses towards moral foundations

Background

- Surveys consist of a questionnaire and human responses.
- They are increasingly used to analyze and evaluate LLMs.
- Can such instruments be applied to LLMs in a sensible manner?

5: Steering LLMs responses towards moral foundations

Background

- Surveys consist of a questionnaire and human responses.
- They are increasingly used to analyze and evaluate LLMs.
- Can such instruments be applied to LLMs in a sensible manner?

Research problem

- You are given a questionnaire and human responses.
- Can we make the LLMs respond to the questionnaire in a human-like way?
- Can we steer their responses using activation and prompt steering?
- How uncertain the models are about their responses?

5: Steering LLMs responses towards moral foundations

Moral Foundations Questionnaire (MFQ-30)

- An instrument measuring five moral foundations: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation [Graham et al., 2011, Graham et al., 2013]
- Human sample from a Norwegian population (N=1282) [Enstad and Finseraas, 2024]
- Can we observe human-like patterns in model responses?
  - E.g., human respondents who score high on care and fairness, tend to score low on loyalty, authority, and sanctity, and vice versa.

5: Steering LLMs responses towards moral foundations

Moral Foundations Questionnaire (MFQ-30)

- An instrument measuring five moral foundations: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation [Graham et al., 2011, Graham et al., 2013]
- Human sample from a Norwegian population (N=1282) [Enstad and Finseraas, 2024]
- Can we observe human-like patterns in model responses?
  - E.g., human respondents who score high on care and fairness, tend to score low on loyalty, authority, and sanctity, and vice versa.

Models to try (non-exhaustive)

- `NorMistral 7B Instruct`
- `NorMistral 11B Thinking`
- `Qwen3 8B`
- `Qwen3 14B`

5: Steering LLMs responses towards moral foundations

Making the models respond to the MFQ-30

- Set up an experimental framework
  - Questionnaire presentation
  - Answer option perturbations
  - Generation strategies etc.
- Compare your results to the human sample

5: Steering LLMs responses towards moral foundations

Making the models respond to the MFQ-30

- Set up an experimental framework
  - Questionnaire presentation
  - Answer option perturbations
  - Generation strategies etc.
- Compare your results to the human sample

Estimating model uncertainty

- Are the models consistent in their 'moral values'?
- Or are they producing very different generations with low probabilities?
- Use methods from uncertainty estimation literature [Fadeeva et al., 2026]

5: Steering LLMs responses towards moral foundations

Making the models respond to the MFQ-30

- Set up an experimental framework
  - Questionnaire presentation
  - Answer option perturbations
  - Generation strategies etc.
- Compare your results to the human sample

Estimating model uncertainty

- Are the models consistent in their 'moral values'?
- Or are they producing very different generations with low probabilities?
- Use methods from uncertainty estimation literature [Fadeeva et al., 2026]

Steering the models' foundations

- Can you push the model to certain moral foundations? [Rimsky et al., 2024]
- **Activation steering** focuses on adjusting the internal activations
- **Prompt steering** focuses on systematically adjusting the prompt

Contents

1. Introduction
2. 1: English – Sámi machine translation
3. 2: Disentangling content and formal reasoning in LLMs
4. 3: Predicting Emotional Valence and Arousal over Time
5. 4: Open-Ended Regional Question Answering with RAG
6. 5: Steering LLMs responses towards moral foundations
7. 6: Machine-generated Code Detection
8. What's next?

6: Machine-generated Code Detection

Main info

Track chairs: Vladislav Mikhailov

Learn more: `https://github.uio.no/in5550/2026/tree/main/exam/generated_code_detection`

6: Machine-generated Code Detection

Background

- Modern generative LMs are widely used on a daily basis for software development: code completion, debugging, documentation, etc.
- However, such models can be misused for harmful purposes, e.g., generating vulnerable code, obfuscated malware, plagiarized solutions, or bypassing security and licensing constraints.

6: Machine-generated Code Detection

Background

- Modern generative LMs are widely used on a daily basis for software development: code completion, debugging, documentation, etc.
- However, such models can be misused for harmful purposes, e.g., generating vulnerable code, obfuscated malware, plagiarized solutions, or bypassing security and licensing constraints.

Machine-generated code detection

- A specialized area focused on mitigating the risks associated with the misuse of code-generating LMs.
- Developing datasets, benchmarks, and detection systems to assist humans in identifying machine-generated code.

6: Machine-generated Code Detection

Goal

- Develop your own system for a task formulation of your choice.

6: Machine-generated Code Detection

Goal

- Develop your own system for a task formulation of your choice.

Data

- SemEval-2026 Task 13: Detecting Machine-Generated Code with Multiple Programming Languages, Generators, and Application Scenarios
- Robustness across different programming languages, domains, and generation techniques.

6: Machine-generated Code Detection

Goal

- Develop your own system for a task formulation of your choice.

Data

- SemEval-2026 Task 13: Detecting Machine-Generated Code with Multiple Programming Languages, Generators, and Application Scenarios
- Robustness across different programming languages, domains, and generation techniques.

Task Formulations

- Subtask A: is the code snippet created by a human or an LM?
- Subtask B: if latter, which LM generated the code?
- Subtask C: is the code snippet multi-authored?

6: Machine-generated Code Detection

General Workflow

- Select one or more task formulations.
- Analyze the participants' submissions, think about how they can be improved or come up with your own solutions.
- Try various types of detectors (zero-shot, feature-based, and Transformer-based detectors) and their ensembles.
- Conduct ablation studies to analyze the performance w.r.t. robustness, training data mixtures, dataset criteria.
- Compare the performance of your resulting system against the corresponding public leaderboards.

Contents

1. Introduction
2. 1: English – Sámi machine translation
3. 2: Disentangling content and formal reasoning in LLMs
4. 3: Predicting Emotional Valence and Arousal over Time
5. 4: Open-Ended Regional Question Answering with RAG
6. 5: Steering LLMs responses towards moral foundations
7. 6: Machine-generated Code Detection
8. What's next?

What's next?

- Before April 30:
  - browse the tracks, choose your favorite
  - decide on your team composition (if any)
  - inform us via e-mail

What's next?

- Before April 30:
  - browse the tracks, choose your favorite
  - decide on your team composition (if any)
  - inform us via e-mail
- Once we know who takes what topic, track chairs will contact you for further planning

What's next?

- Before April 30:
  - browse the tracks, choose your favorite
  - decide on your team composition (if any)
  - inform us via e-mail
- Once we know who takes what topic, track chairs will contact you for further planning
- Make sure you have an OpenReview account.
- Submit your research paper by May 20.

What's next?

- Before April 30:
  - browse the tracks, choose your favorite
  - decide on your team composition (if any)
  - inform us via e-mail
- Once we know who takes what topic, track chairs will contact you for further planning
- Make sure you have an OpenReview account.
- Submit your research paper by May 20.
- And don't forget the guest industry lecture next Monday!
  - **'Neural methods in online grocery shopping (at Oda)'** by Fredrik Jørgensen and Lars Mushom (in person)
- No other lectures or labs during the exam period

What's next?

- Before April 30:
  - browse the tracks, choose your favorite
  - decide on your team composition (if any)
  - inform us via e-mail
- Once we know who takes what topic, track chairs will contact you for further planning
- Make sure you have an OpenReview account.
- Submit your research paper by May 20.
- And don't forget the guest industry lecture next Monday!
  - **'Neural methods in online grocery shopping (at Oda)'** by Fredrik Jørgensen and Lars Mushom (in person)
- No other lectures or labs during the exam period

Best of luck, see you at the WNNLP 2026 workshop!

References I

Enstad, J. D. and Finseraas, H. (2024).
Moralske intuisjoner og politiske orienteringer blant norske velgere.
Tidsskrift for samfunnsforskning, 65(1):1–23.

Fadeeva, E., Goloburda, M., Rubashevskii, A., Vashurin, R., Shelmanov, A., Nakov, P., Sachan, M., and Panov, M. (2026).
Don't throw away your beams: Improving consistency-based uncertainties in LLMs via beam search.
In The Fourteenth International Conference on Learning Representations.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. (2013).
Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism.

References II

In Devine, P. and Plant, A., editors, Advances in Experimental Social Psychology, volume 47, pages 55–130. Academic Press.



Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., and Ditto, P. H. (2011).
Mapping the Moral Domain.
Journal of Personality and Social Psychology, 101(2):366–385.



Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. (2024).
Steering llama 2 via contrastive activation addition.
In Ku, L.-W., Martins, A., and Srikumar, V., editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.