KORMo: Korean Open Reasoning Model for Everyone

Minjun Kim[1,4*]   **Hyeonseok Lim**[1,4*]   **Hangyeol Yoo**[1,4*]   **Inho Won**[1*]
Seungwoo Song[1,4]   **Minkyung Cho**[1]   **Junghun Yuk**[1]   **Changsu Choi**[1,4]
Dongjae Shin[1,4]   **Huije Lee**[2]   **Hoyun Song**[1]
Alice Oh[3]   **KyungTae Lim**[1]

[1]KAIST MLP Lab   [2]KAIST NLPCL Lab   [3]KAIST U&I Lab   [4]SeoulTech
Corresponding Author: ktlim@kaist.ac.kr

- Problem / objective
  - Synthetic data-driven Fully Open Model (FOM) for Korean

- Contribution / Key idea
  - **KORMo-10B**: The first Korean-English fully-open LLM with 10.8B parameter
    i. Synthetic data: 68.73% of the Korean corpus
       - Synthetic data with careful linguistic balance and stylistic diversity can replace large-scale human-curated data
    ii. Training curriculum totaling 150 billion tokens

전유진