

1. 실험에 사용한 문장 예시

- “나는 늦게 성장하기 시작한 late-bloomer야. 그래서 난 희망을 가져.”

2. 사용한 tokenizer 이름 및 링크

- DeepSeek-V3-0324
(<https://huggingface.co/deepseek-ai/DeepSeek-V3-0324>)
- openchat/openchat-3.5-0106
(<https://huggingface.co/openchat/openchat-3.5-0106>)

3. 각 tokenizer에서 분해된 token 수 및 token 목록

- openchat-3.5-0106 결과

입력 문장: 나는 늦게 성장하기 시작한 late-bloomer야. 그래서 난 희망을 가져.

Token 개수: 41

Token 리스트: ['_', '나', '는', ' ', '<0xEB>', '<0x8A>', '<0xA6>', '게', ' ', '성', '장', '하', '기', ' ', '시', '작', '한', ' ', '_late', '-', 'b1 o', 'omer', '야', ' ', ' ', '그', '래', '서', ' ', '난', ' ', ' ', '<0xED>', '<0x9D>', '<0xAC>', '<0xEB>', '<0xA7>', '<0x9D>', '을', ' ', '가', '저', ' ', ':']

Input IDs: [28705, 29695, 29175, 28705, 238, 141, 169, 29833, 28705, 29465, 29747, 29136, 29164, 28705, 29236, 29781, 29282, 3909, 28733, 15190, 16245, 30263, 28723, 28705, 29614, 29889, 29305, 28705, 31876, 28705, 240, 160, 175, 238, 170, 160, 29189, 28705, 29135, 30942, 28723]

- Deepseek-V3-0324 결과

입력 문장: 나는 늦게 성장하기 시작한 late-bloomer야. 그래서 난 희망을 가져.

12 Token 개수: 76
34

Token 리스트: ['Ė', 'Ĥ', 'Ĭ', 'ē', 'İ̇', 'K̇', 'Ġ', 'ē', 'İ̇', '|', 'ē', '²', 'Ĭ', 'Ġ', 'ì', 'Ḣ', '±', 'ì', 'l̇', '¶', 'ı́', 'k̇', 'Ĭ', 'ē', '.', '°', 'Ġ', 'ì', 'ı́', 'İ', 'ı́', 'l̇', 'j̇', 'ı́', 'k̇', 'İ', 'Ġlate', '-', 'b', 'loom', 'er', 'ı́', 'k̇', '¼', '.', 'Ġ', 'ē', '.', '.', 'ē', 'l̇', 'Ĭ', 'ı́', 'Ḣ', 'İ', 'Ġ', 'ē', 'Ĥ', 'İ', 'Ġ', 'ı́', 'l̇', 'ı́', 'ē', 'Ġ', 'l̇', 'il̇', 'Ḣ', 'Ġ', 'ē', '°', 'Ġ', 'ı́', 'İ', '.', '.']

Input IDs: [165, 211, 233, 165, 219, 229, 207, 165, 219, 99, 164, 110, 221, 207, 166, 213, 109, 166, 239, 98, 167, 230, 233, 164, 116, 108, 207, 16
6, 220, 237, 166, 239, 226, 167, 230, 237, 5179, 12, 65, 13133, 250, 166, 230, 120, 13, 207, 164, 115, 116, 165, 239, 233, 166, 213, 237, 207, 165, 21
1, 237, 207, 167, 238, 105, 165, 100, 238, 24625, 213, 207, 164, 108, 209, 166, 241, 116, 13]

4. 가장 흥미로웠던 token 분해 결과와 이유

- OpenChat은 유니코드 방식으로 분해한 것 같은데, 딥시크는 토큰이 많은 대신 그걸 효율적으로 잘게 잘라 메모리를 경량화한게 흥미로웠습니다. 두 토큰나이저의 접근 방식이 확연히 다른 것을 볼 수 있었습니다.

OpenChat-3.5는 한글 문자를 비교적 온전하게 보존하며 '나', '는', '늦', '게'와 같이 음절 단위로 토큰화했고, 영어 단어인 'late-bloomer'도 음절 단위로 유지했습니다. 반면 Deepseek-V3는 더 세밀한 수준으로 토큰을 분해하여 전체적으로 토큰 수는 많아졌지만, 각 토큰의 크기를 줄여 메모리 사용을 최적화한 것으로 보입니다. 이러한 차이는 각 모델의 설계 철학과 자원 활용 전략을 반영하는 것으로, 특히 Deepseek-V3의 경량화 접근 방식이 주목할 만했습니다.

앞으로는 주제를 더 정확히 이해하고 분석하여 보고서를 작성하겠습니다. 이번에는 두 문장 이상을 tokenizer에다가 하라는 권고사항을 읽지 못하고 성급하게 실험했던 점을 인정합니다.