Upstage, which has gained attention from the global AI industry for building the world's best open LLM model, is launching the 'Korean LLM Independence Declaration'.

Upstage (CEO Kim Sung-hoon) announced on the 14th that it will launch the '1T Club' to solve the problem of lack of Korean data and achieve independence of Korean LLM through the development of high-performance LLM (Large Language Model: giant language model). The '1T Club' is a shortened version of the '1 Trillion Token Club', which consists of partner companies that contribute more than 100 million words of Korean data in various forms such as text, books, articles, reports, and papers.

Upstage has recently attracted attention by surpassing the performance of GPT-3.5, the basis of ChatGPT, and ranking first in the evaluation score of the 'Open LLM Leaderboard' operated by Hugging Face, the world's largest machine learning platform.

The Hugging Face Open LLM Leaderboard is considered a barometer of open-source generative AI models. More than 500 open models around the world compete in the average score of four indicators, including inference and common sense ability, language understanding comprehensive ability, and hallucination prevention, and a credible ranking is made. Currently, Upstage has deployed a model that exceeds 73 points, monopolizing the world's first and second place models.

The '1T Club' is a new challenge by Upstage, which has been struggling to solve the problem of lack of Korean data and establish an ecosystem in which data providers and model production companies coexist for the independence of domestic LLM. Upstage expects to be able to develop high-quality LLM that can contain Korean cultural emotions and contribute to the development of artificial intelligence by securing and sharing Korean data through the 1T Club and utilizing it in various applications in the domestic generative AI field.

Upstage is currently in close consultation with more than 20 media companies, companies, and academia for partnerships. In addition, it plans to promote cooperation with various industry-leading companies to build private LLM, as well as with partner companies in various fields that will contribute to the development of Korean LLM. Those interested in the '1T Club' can submit an application through the Upstage official website or link (www.upstage.ai/up-1-trillion-token-club).

Korean data is an essential resource for the development of Korean LLM, but it is currently quite scarce and is facing copyright issues. LLMs of big tech companies trained in foreign languages are vulnerable to Korean language skills, emotions, and local information, which is a stumbling block to the development of private LLMs that domestic companies will use.

For example, in the case of Meta's 'Llama2', which is considered the best model in the open LLM market recently, it was trained with 2 trillion tokens, and Google's 'Lambda' was trained with 2.81 trillion tokens, showing amazing performance. However, the amount of Korean data training is about 100 million tokens based on GPT-3, accounting for only 0.01697% of the total and ranking 28th among all languages. However, English was trained with 45 trillion tokens, creating a gap in the performance of LLM by language proportional to the amount of data.

Upstage plans to do its best to further enhance Korea's AI capabilities and establish Korea's position as a leading player in the global AI industry through the '1T Club'. In particular, it plans to operate so that both data providers and model producers can benefit, as well as solve the problem of side effects such as copyright issues due to AI learning through crawling.

Upstage plans to provide benefits to partner companies participating in the '1T Club' in two ways: discounting API usage fees in proportion to the amount of data provided, and sharing profits generated by the LLM's API business.

First, in the case of the former, API usage fee discounts, partner companies can use the API of the highest performance LLM produced by Upstage at a discounted price in proportion to the number of tokens contributed, and use it for various applications. For example, a partner company that has provided 100 million words of tokens can use the API equivalent to 100 million tokens free of charge.

In addition, the 'Profit Share method', which shares profits, is operated by sharing a portion of the profits generated by the LLM's API business with partner companies when Upstage generates profits from the LLM's API business. Upstage plans to use a portion of the LLM API business revenue as a source of funds for this purpose and allocate it to the 1T Club profit sharing, and each partner company can receive profits in proportion to the amount of data it has contributed.

Upstage plans to make every effort to ensure the security and privacy of the provided data. Upstage plans to use the data provided by partner companies only for the purpose of pre-training Korean in the model, so that it will only have the ability to summarize and organize general knowledge and articles, and will not be able to extract the original text. In addition, it plans to use its own jailbreak check technology to prevent the original text from being leaked to the outside or used for other purposes.

Upstage CEO Kim Sung-hoon said, "LLM is the core technology of generative artificial intelligence today, and it is important to create an ecosystem so that companies in various industries in Korea can freely utilize high-performance private LLM." He added, "We will do our best to protect the rights of data providers through the '1T Club' and develop LLM that can contain Korean cultural emotions based on this, so that all companies in Korea can benefit from the development of AI."